NEW AGE

RESEARCH METHODOLOGY

METHODS AND TECHNIQUES

FOURTH
Multi Colour
EDITION

Primary Research
Client Internal Data
In-house Repository
Data Validation
360° Research
Data Analysis
Data Mining
Social Media Research
Secondary Research
Competitors Sources

C R KOTHARI • GAURAV GARG

NEW AGE INTERNATIONAL PUBLISHERS

NATIONAL LAW SCHOOL
OF INDIA UNIVERSITY
BENGALURU

Please remember that this resource is meant for many other users like you, so do not tamper, tear or damage.

We believe that it is an individual responsibility to care for library documents and other resources of institution.

Please do check before you borrow, because you have to compensate if the material is returned in a damaged condition.

RESEARCH METHODOLOGY

METHODS AND TECHNIQUES

RESEARCH METHODOLOGY

METHODS AND TECHNIQUES

FOURTH
Multi Colour
EDITION

C R KOTHARI
Former Principal, College of Commerce
University of Rajasthan, Jaipur
Rajasthan (India)

GAURAV GARG
Assistant Professor
Indian Institute of Management
Lucknow, Uttar Pradesh (India)

NEW AGE
TECHNO
PRESS

An Imprint of

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
LONDON • NEW DELHI • NAIROBI
Bangalore • Chennai • Cochin • Guwahati • Hyderabad • Kolkata • Lucknow • Mumbai

Visit us at **www.newagepublishers.com**

Copyright © 2019, 2014, 2004, 1990, 1985, New Age International (P) Ltd., Publishers

Published by New Age International (P) Ltd., Publishers
First Edition: 1985
Fourth Multi Colour Edition: 2019
Reprint: 2022

All rights reserved.
No part of this book may be reproduced in any form, by photostat, microfilm, xerography, or any other means, or incorporated into any information retrieval system, electronic or mechanical, without the written permission of the publisher.

GLOBAL OFFICES

- **New Delhi** NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
  7/30 A, Daryaganj, New Delhi-110002, (INDIA)
  Tel.: (011) 23253771, 23253472, **Telefax:** 23267437, 43551305
  E-mail: contactus@newagepublishers.com • **Visit us at** www.newagepublishers.com
- **London** NEW AGE INTERNATIONAL (UK) LTD.
  27 Old Gloucester Street, London, WC1N 3AX, UK
  E-mail: info@newacademicscience.co.uk • **Visit us at** www.newacademicscience.co.uk
- **Nairobi** NEW AGE GOLDEN (EAST AFRICA) LTD.
  Ground Floor, Westlands Arcade, Chiromo Road (Next to Naivas Supermarket)
  Westlands, Nairobi, KENYA, **Tel.:** 00-254-713848772, 00-254-725700286
  E-mail: kenya@newagepublishers.com

BRANCHES

- **Bangalore** 37/10, 8th Cross (Near Hanuman Temple), Azad Nagar, Chamarajpet, Bangalore- 560 018
  Tel.: (080) 26756823, **Telefax:** 26756820, **E-mail:** bangalore@newagepublishers.com
- **Chennai** 26, Damodaran Street, T. Nagar, Chennai-600 017, **Tel.:** (044) 24353401, **Telefax:** 24351463
  E-mail: chennai@newagepublishers.com
- **Cochin** CC-39/1016, Carrier Station Road, Ernakulam South, Cochin-682 016, **Tel.:** (0484) 2377303
  Telefax: 4051304
  E-mail: cochin@newagepublishers.com
- **Guwahati** Hemsen Complex, Mohd. Shah Road, Paltan Bazar, Near Starline Hotel, Guwahati-781 008,
  Tel.: (0361) 2513881, **Telefax:** 2543669, **E-mail:** guwahati@newagepublishers.com
- **Hyderabad** 105, 1st Floor, Madhiray Kaveri Tower, 3-2-19, Azam Jahi Road, Near Kumar Theater, Nimboliadda
  Kachiguda, Hyderabad-500 027, **Tel.:** (040) 24652456, **Telefax:** 24652457
  E-mail: hyderabad@newagepublishers.com
- **Kolkata** RDB Chambers (Formerly Lotus Cinema) 106A, 1st Floor, S N Banerjee Road, Kolkata-700 014
  Tel.: (033) 22273773, **Telefax:** 22275247, **E-mail:** kolkata@newagepublishers.com
- **Lucknow** 16-A, Jopling Road, Lucknow-226 001, **Tel.:** (0522) 2209578, 4045297, **Telefax:** 2204098
  E-mail: lucknow@newagepublishers.com
- **Mumbai** 142C, Victor House, Ground Floor, N.M. Joshi Marg, Lower Parel, Mumbai-400 013
  Tel.: (022) 24927869, **Telefax:** 24915415, **E-mail:** mumbai@newagepublishers.com
- **New Delhi** 22, Golden House, Daryaganj, New Delhi-110 002, **Tel.:** (011) 23262368, 23262370
  Telefax: 43551305, **E-mail:** sales@newagepublishers.com

ISBN: 978-93-86649-22-5
C-21-10-12847
Printed in India at Goyal Offset Works (P) Ltd., Kundli, Haryana.
Typeset at In-house, Delhi.

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
7/30 A, Daryaganj, New Delhi-110002
Visit us at **www.newagepublishers.com**
(CIN: U74899DL1966PTC004618)

Purchased: MLU
Approval : 038805
Gratis :
Acc No :
Price :
"Narayan Rao Melgiri
National Law Library"
Bangalore
h39a

In loving memory of
my revered father
(The fountain of inspiration)

— C.R. KOTHARI

To my Family

— GAURAV GARG

Preface to the Fourth Edition

Exhaustive and rigorous transformation in third edition from second edition left very less scope of further improvement in this fourth edition. However, as improvement is a continuous and never ending process, we present this fourth edition to you. In this edition, we corrected some mistakes and added some new problems along with the datasets. No new topic has been added and no change has been made in overall structure of the book. The datasets used in the solved and unsolved problems are available on the New Age International Publishers website:

(*i*) Go to www.newagepublishers.com
(*ii*) On the left top, go to search and enter author name or title name or ISBN number
(*iii*) Book details appear. Open the book page
(*iv*) Under 'About the Book' click at supplementary data.

It has been great working with New Age International Publishers. It is good to see that they believe in continuous improvement of this popular textbook. I look forward for your feedback based on which we may think of improving this text further.

GAURAV GARG

Preface to the Third Edition

I feel privileged to get associated with this popular book as co-author. This third edition is presented after extensive restructuring and with more details. The revised edition is prepared keeping in mind the need of advanced data analysis in social science research.

In this edition, old chapters are restructured to make the learning more systematic. Some details, questions, and solved examples are added to old chapters wherever needed. Some old chapters are renamed, some merged, and some split. Some new chapters are also added in this edition. These added chapters are mainly on advanced level data analysis. The usage of statistical package SPSS has also been shown through many examples.

For ancillary matter on this book, click at the following link:
http://www.newagepublishers.com/servlet/nagetbiblio?bno=000896

I am thankful to New Age International Publishers for giving me the opportunity of getting associated with this book. I also look forward for your suggestions and feedbacks so that the text can be improved further.

GAURAV GARG

Preface to the Second Edition

I feel encouraged by the widespread response from teachers and students alike to the first edition. I am presenting this second edition, thoroughly revised and enlarged, to my readers in all humbleness. All possible efforts have been made to enhance further the usefulness of the book. The feedback received from different sources has been incorporated.

In this edition a new chapter on *"The Computer: Its role in Research"* have been added in view of the fact that electronic computers by now, for students of economics, management and other social sciences, constitute an indispensable part of research equipment.

The other highlights of this revised edition are (i) the subject contents has been developed, refined and restructured at several points, (ii) several new problems have also been added at the end of various chapters for the benefit of students, and (iii) every page of the book has been read very carefully so as to improve its quality.

I am grateful to all those who have helped me directly and/or indirectly in preparing this revised edition. I firmly believe that there is always scope for improvement and accordingly I shall look forward to received suggestions, (which shall be thankfully acknowledged) for further enriching the quality of the next.

C.R. KOTHARI

Preface to the First Edition

Quite frequently these days people talk of research, both in academic institutions and outside. Several research studies are undertaken and accomplished year after year. But in most cases very little attention is paid to an important dimension relating to research, namely, that of research methodology. The result is that much of research, particularly in social sciences, contains endless word-spinning and too many quotations. Thus a great deal of research tends to be futile. It may be noted, in the context of planning and development, that the significance of research lies in its quality and not in quantity. The need, therefore, is for those concerned with research to pay due attention to designing and adhering to the appropriate methodology throughout for improving the quality of research. The methodology may differ from problem to problem, yet the basic approach towards research remains the same.

Keeping all this in view, the present book has been written with two clear objectives, viz., (i) to enable researches, irrespective of their discipline, in developing the most appropriate methodology for their research studies; and (ii) to make them familiar with the art of using different research-methods and techniques. It is hoped that the humble effort made in the form of this book will assist in the accomplishment of exploratory as well as result-oriented research studies.

Regarding the organization, the book consists of fourteen chapters, well arranged in coherent manner. Chapter One is an introduction, presenting an overview of the research methodology. Chapter Two explains the technique of defining a research problem. Chapter Three dwells on various research designs, highlighting their main characteristics. Chapter Four presents the details of several sampling designs. Different measurement and scaling techniques, along with multidimensional scaling, have been lucidly described in Chapter Five. Chapter Six presents a comparative study of the different methods of data collection. It also provides in its appendices guidelines for successful interviewing as well as for constructing questionnaire/schedules. Chapter Seven deals with processing and analysis of data. Sampling fundamentals, along with the theory of estimation, constitutes the subject-matter of Chapter Eight. Chapter Nine has been exclusively devoted to several parametric tests of hypotheses, followed by Chapter Ten concerning Chi-square test. In Chapter Eleven important features of ANOVA and ANOCOVA techniques have been explained and illustrated. Important non-parametric tests, generally used by researchers have been described and illustrated in Chapter Twelve. In Chapter Thirteen, an effort has been made to present the conceptual aspects and circumstances under which various multivariate techniques can appropriate be utilized in research studies, specially in behavioural and social sciences. Factor analysis has been dealt with in relatively more detail. Chapter Fourteen has been devoted to the task of interpretation and the art of writing research reports.

The book is primarily intended to serve as a textbook for graduate and M.Phil. students of Research Methodology in all disciplines of various universities. It is hoped that the book shall provide guidelines to all interested in research studies of one sort or the other. The book is, in fact, an outgrowth of my experience of teaching the subject to M.Phil. students for the last several years.

I am highly indebted to my students and learned colleagues in the Department for providing the necessary stimulus for writing this book. I am grateful to all those persons whose writings and works have helped me in the preparation of this book. I am equally grateful to the reviewer of the manuscript of this book who made extremely valuable suggestions and has thus contributed in enhancing the standard of the book. I thankfully acknowledge the assistance provided by the University Grants Commission in the form of 'on account' grant in the preparation of the manuscript of this book.

I shall feel amply rewarded if the book proves helpful in the development of genuine research studies. I look forward to suggestions from all readers, specially from experienced researchers and scholars for further improving the subject content as well as the presentation of this book.

C.R. KOTHARI

Contents

Preface to the Fourth Edition (vii)
Preface to the Third Edition (ix)
Preface to the Second Edition (x)
Preface to the First Edition (xi)

1. Research Methodology: An Introduction **1**

1.1 Meaning of Research 1
1.2 Objectives of Research 2
1.3 Types of Research 2
1.4 Research Approaches 4
1.5 Significance of Research 5
1.6 Research Methods versus Methodology 6
1.7 Research and Scientific Method 8
1.8 Research Process 9
1.9 Criteria of Good Research 19
1.10 Problems Encountered by Researchers in India 20
- *Problems* 21

2. Defining the Research Problem **22**

2.1 What is a Research Problem? 22
2.2 Selecting the Problem 23
2.3 Necessity of Defining the Problem 24
2.4 Technique Involved in Defining a Problem 25

2.5 An Illustration 27
2.6 Conclusion 2
- *Problems* 2

3. Research Design 29

3.1 Meaning of Research Design 29
3.2 Need for Research Design 30
3.3 Features of a Good Design 3
3.4 Important Concepts Relating to Research Design 3
  3.4.1 Dependent and Independent Variables 32
  3.4.2 Extraneous Variable 32
  3.4.3 Control 32
  3.4.4 Confounded Relationship 32
  3.4.5 Research Hypothesis 33
  3.4.6 Experimental and Non-Experimental Hypothesis-Testing Research 33
  3.4.7 Experimental and Control Groups 33
  3.4.8 Treatments 33
  3.4.9 Experiment 34
  3.4.10 Experimental Unit(s) 34
3.5 Different Research Designs 34
  3.5.1 Research Design in Case of Exploratory Research Studies 34
  3.5.2 Research Design in Case of Descriptive and Diagnostic Research Studies 35
  3.5.3 Research Design in Case of Hypothesis-Testing Research Studies 38
3.6 Basic Principles of Experimental Designs 38
3.7 Important Experimental Designs 39
  3.7.1 Before-and-After Without Control Design 40
  3.7.2 After-Only with Control Design 40
  3.7.3 Before-and-After with Control Design 41
  3.7.4 Completely Randomized Design (C.R. Design) 41
  3.7.5 Randomized Block Design (R.B. Design) 44
  3.7.6 Latin Square Design (L.S. Design) 45
  3.7.7 Factorial Designs 46
3.8 Conclusion 51
- *Problems* 51

4. Design of Sample Surveys 52

4.1 Introduction 52
4.2 Sample Design 52

4.3 Sampling and Non-sampling Errors 54
    4.3.1 Sampling Errors 54
    4.3.2 Non-sampling Errors 55
4.4 Sample Survey Vs. Census Survey 55
4.5 Types of Sampling Designs 55
    4.5.1 Non-probability Sampling 55
    4.5.2 Probability Sampling 56
    4.5.3 Complex Random Sampling Designs 58
    - *Problems* 64

5. Measurement and Scaling **66**

5.1 Quantitative and Qualitative Data 66
5.2 Classifications of Measurement Scales 67
5.3 Goodness of Measurement Scales 70
5.4 Sources of Error in Measurement 72
5.5 Techniques of Developing Measurement Tools 72
5.6 Scaling 73
5.7 Scale Classification Bases 74
5.8 Scaling Techniques 75
    5.8.1 Comparative Scaling Techniques 76
    5.8.2 Non-comparative Scaling Techniques 79
5.9 Multidimensional Scaling 86
5.10 Deciding the Scale 87
    - *Summary* 87
    - *Problems* 87

6. Data Collection **89**

6.1 Introduction 89
6.2 Experiments and Surveys 89
6.3 Collection of Primary Data 91
    6.3.1 Difference between Questionnaire and Schedule 100
    6.3.2 Guidelines for Constructing Questionnaire/Schedule 101
    6.3.3 Some Other Methods of Data Collection 102
6.4 Collection of Secondary Data 107
6.5 Selection of Appropriate Method for Data Collection 108
6.6 Case Study Method 109
    - *Problems* 113

7. Data Preparation **114**

7.1 Data Preparation Process 114
- 7.1.1 Questionnaire Checking 114
- 7.1.2 Editing 115
- 7.1.3 Coding 115
- 7.1.4 Classification 116
- 7.1.5 Tabulation 119
- 7.1.6 Graphical Representation 121
- 7.1.7 Data Cleaning 124
- 7.1.8 Data Adjusting 124

7.2 Some Problems in Preparation Process 125

7.3 Missing Values and Outliers 126

7.4 Types of Analysis 126

7.5 Statistics in Research 127
- *Problems* 128

8. Descriptive Statistics **129**

8.1 Measures of Central Tendency 129
- 8.1.1 Mean 129
- 8.1.2 Median 130
- 8.1.3 Mode 131
- 8.1.4 Other Averages 131

8.2 Measures of Dispersion 132
- 8.2.1 Range 132
- 8.2.2 Mean Deviation 132
- 8.2.3 Standard Deviation 133

8.3 Measures of Skewness 134

8.4 Kurtosis 136

8.5 Measures of Relationship 136
- 8.5.1 Covariance 138
- 8.5.2 Karl Pearson's Coefficient of Correlation 139
- 8.5.3 Rank Correlation 140

8.6 Association in Case of Attributes 140

8.7 Other Measures 143
- 8.7.1 Index Numbers 143
- 8.7.2 Time Series 144
- *Problems* 146

9. Sampling and Statistical Inference 147
9.1 Parameter and Statistic 148
9.2 Sampling and Non-sampling Errors 148
9.3 Sampling Distribution 149
9.3.1 Sampling Distribution of Mean 150
9.3.2 Sampling Distribution of Proportion 150
9.3.3 Student's t-Distribution 151
9.3.4 Chi-square (χ^2) Distribution 151
9.3.5 Snedecor's F-Distribution 152
9.4 Degree of Freedom 152
9.5 Standard Error 152
9.6 Central Limit Theorem 153
9.7 Finite Population Correction 154
9.8 Statistical Inference 154
9.8.1 Point Estimation 154
9.8.2 Interval Estimation 156
9.8.3 Sample Size and its Determination 162
9.8.4 Tests of Significance (Hypothesis Testing) 170
- *Problems* 175

10. Testing of Hypothesis 179
10.1 What is a Hypothesis? 179
10.2 Basic Concepts Concerning Testing of Hypothesis 180
10.2.1 Null Hypothesis and Alternative Hypothesis 180
10.2.2 Type I and Type II Errors 181
10.2.3 Level of Significance 182
10.2.4 Two-Tailed and One-Tailed Tests 182
10.3 Testing the Hypothesis 183
10.4 Test Statistic and Critical Region 183
10.5 Critical Value and Decision Rule 184
10.6 Procedure for Hypothesis Testing 185
10.7 Hypothesis Testing for Mean 187
10.8 Hypothesis Testing for Proportion 190
10.9 Hypothesis Testing for Variance 191
10.10 Hypothesis Testing for Difference of Two Mean 194
10.11 Hypothesis Testing for Difference of Two Proportions 198
10.12 Hypothesis Testing for Difference of Two Variances 199

10.13 *P*-Value Approach
10.14 Power of the Test
10.15 Limitations of the Tests of Hypotheses
- *Problems*

11. Chi-Square Tests

11.1 Test of Difference of more than Two Proportions
11.2 Test of Independence of Attributes
  11.2.1 Alternative Formula
  11.2.2 Yates' Correction
  11.2.3 Magnitude of Association
  11.2.4 Coefficient of Contingency
11.3 Test of Goodness of Fit
  11.3.1 Goodness of Fit Test for Normal Distribution
11.4 Caution in Using Chi Square Tests
- *Problems*

12. Analysis of Variance

12.1 The ANOVA Technique
12.2 The Basic Principle of ANOVA
12.3 One Way ANOVA
  12.3.1 Analysis of Variance Table
  12.3.2 Short-cut Method for One-way ANOVA
  12.3.3 Critical Difference
12.4 Two Way ANOVA
  12.4.1 One Observation per Cell
  12.4.2 More than One Observations Per Cell
12.5 Latin-square Design
12.6 Analysis of Co-variance (ANOCOVA)
12.7 Assumptions in ANOCOVA
- *Problems*

13. Other Nonparametric Methods

13.1 Sign Tests
  13.1.1 One Sample Sign Test
  13.1.2 Two Samples Sign Test (Paired Sign Test)

13.2 Wilcoxon Signed Rank Sum Test for Single Population 294
  13.2.1 Wilcoxon Signed Rank Sum Test (Single Sample) 294
  13.2.2 Wilcoxon Signed Rank Sum Test (Paired Samples) or Wilcoxon Matched Pairs Test 295
13.3 Mann Whitney U Test 297
13.4 Run Test 301
13.5 Kruskal Wallis Test 302
13.6 Spearman's Rank Correlation 304
13.7 Some Features of Non-parametric Tests 309
- *Problems* 310

14. Linear Regression Analysis **312**

14.1 Dependent and Independent Variables 312
14.2 Simple Linear Regression Model 313
  14.2.1 Least Squares Estimation 313
  14.2.2 Coefficient of Determination 316
  14.2.3 Standard Error 319
  14.2.4 Assumptions or Conditions Required 319
  14.2.5 Testing the Hypothesis about the Slope and Correlation Coefficient 323
  14.2.6 Confidence Interval Estimate of the Slope 325
  14.2.7 Confidence Interval 326
  14.2.8 Prediction Interval 326
  14.2.9 Using Excel 327
14.3 Multiple Linear Regression Model 328
  14.3.1 Least Squares Estimation 329
  14.3.2 R^2 and Adjusted R^2 Coefficients 331
  14.3.3 Standard Error 331
  14.3.4 Assumptions 332
  14.3.5 *t*-Test for the Slopes 333
  14.3.6 *F*-Test (ANOVA) 334
  14.3.7 Confidence Interval Estimates of the Regression Coefficients 335
  14.3.8 Confidence Interval of Mean Response and Prediction Interval of Individual Response 336
14.4 Problem of Multicolinearity 336
  14.4.1 Variance Inflationary Factor 339
  14.4.2 Variable Elimination 340

14.5 Qualitative Explanatory Variables 3
14.6 Using SPSS 3
- *Problems* 3

15. Factor Analysis 35

15.1 The Mathematical Basis 3
15.2 Important Methods of Factor Analysis 3
  15.2.1 Centroid Method 3
  15.2.2 Principal Components Method 3
  15.2.3 Maximum Likelihood Method 3
15.3 Rotation in Factor Analysis 3
15.4 *R*-Type and *Q*-Type Factor Analysis 3
15.5 Merits and Demerits of Factor Analysis 3
15.6 Using SPSS 3
- *Problems* 3

16. Discriminant Analysis 38

16.1 Introduction 3
16.2 Two Group Discriminant Analysis 3
  16.2.1 Methodology 3
  16.2.2 Assumptions 3
  16.2.3 Using SPSS 3
  16.2.4 Predicting Group Membership 3
16.3 Multiple Discriminant Analysis 3
  16.3.1 Methodology 3
  16.3.2 Using SPSS 3
- *Problems* 3

17. Cluster Analysis 39

17.1 Introduction 3
17.2 Distance Measures 3
17.3 Clustering Algorithms 3
  17.3.1 Non-hierarchical Clustering 3
  17.3.2 Hierarchical Clustering 4
17.4 Agglomerative Clustering 4
  17.4.1 Combining Clusters 4
  17.4.2 Using SPSS 4
- *Problems* 4

18. Other Multivariate Techniques 409

18.1 Characteristics and Applications 410
18.2 Classification of Multivariate Techniques 410
18.3 Some Multivariate Techniques 411
    18.3.1 Path Analysis 412
    18.3.2 Canonical Correlation 413
    18.3.3 Multidimensional Scaling 414
    18.3.4 Multivariate ANOVA 414
    18.3.5 Latent Structure Analysis 415
18.4 Conclusions 415
- *Problems* 415

19. Interpretation and Report Writing 416

19.1 Meaning of Interpretation 416
19.2 Techniques of Interpretation 417
19.3 Precautions in Interpretation 417
19.4 Significance of Report Writing 418
19.5 Different Steps in Writing Report 419
19.6 Layout of the Research Report 420
19.7 Types of Reports 422
19.8 Oral Presentation 425
19.9 Mechanics of Writing a Research Report 425
19.10 Precautions for Writing Research Reports 430
19.11 Conclusion 431
- *Problems* 431

Appendix 433

Selected References and Recommended Readings 449

Index 455

1

Chapter

Research Methodology: An Introduction

1.1 MEANING OF RESEARCH

Research in common parlance refers to a search for knowledge. One can also define research as a scientific and systematic search for pertinent information on a specific topic. In fact, research is an art of scientific investigation. Dictionary definition of research is a careful investigation or inquiry specially through search for new facts in any branch of knowledge. Some people consider research as a movement from the known to the unknown. It is actually a voyage of discovery. We all possess the vital instinct of inquisitiveness. When the unknown confronts us, more and more our inquisitiveness makes us probe and attain understanding of the unknown. This inquisitiveness is the mother of all knowledge and the method, which one employs for obtaining the knowledge of whatever the unknown, can be termed as research.

Research is an academic activity and as such the term should be used in a technical sense. According to Clifford Woody, research comprises defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organising and evaluating data; making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis. D. Slesinger and M. Stephenson in the Encyclopaedia of Social Sciences define research as "the manipulation of things, concepts or symbols for the purpose of generalising to extend, correct or verify knowledge, whether that knowledge aids in construction of theory or in the practice of an art." Research is, thus, an original contribution to the existing stock of knowledge making for its advancement. It is the pursuit of truth with the help of study, observation, comparison and experiment. In short, the search for knowledge through objective and systematic method of finding solution to a problem is research. The systematic approach concerning generalisation and the formulation of a theory is also research. As such the term 'research' refers to the systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analysing the facts and reaching certain conclusions either in the form of solution(s) towards the concerned problem or in certain generalisations for some theoretical formulation.

1.2 OBJECTIVES OF RESEARCH

The purpose of research is to discover answers to questions through the application of scienti procedures. The main aim of research is to find out the truth which is hidden and which has not be discovered as yet. Though each research study has its own specific purpose, we mention so general objectives of research below:

(i) To gain familiarity with a phenomenon or to achieve new insights into it (studies with t object in view are termed as *exploratory* or *formulative* research studies);

(ii) To portray accurately the characteristics of a particular individual, situation or a gro (studies with this object in view are known as *descriptive* research studies);

(iii) To determine the frequency with which something occurs or with which it is associa with something else (studies with this object in view are known as *diagnostic* resea studies);

(iv) To test a hypothesis of a causal relationship between variables (such studies are known *hypothesis-testing* research studies).

1.3 TYPES OF RESEARCH

The basic types of research are as follows:

(i) *Descriptive vs. Analytical: Descriptive research* includes surveys and fact-finding enqui of different kinds. The major purpose of descriptive research is description of the state affairs as it exists at present. In social science and business research we quite often the term *Ex post facto research* for descriptive research studies. The main characteris of this method is that the researcher has no control over the variables; he can only rep what has happened or what is happening. Most *ex post facto research* projects are u for descriptive studies in which the researcher seeks to measure such items as, for examp frequency of shopping, preferences of people, or similar data. *Ex post facto studies* a include attempts by researchers to discover causes even when they cannot control variables. The methods of research utilized in descriptive research are survey methods all kinds, including comparative and correlational methods. In *analytical research*, on other hand, the researcher has to use facts or information already available, and anal these to make a critical evaluation of the material.

(ii) *Applied vs. Fundamental: Applied research* aims at finding a solution for an immedi problem facing a society or an industrial/business organisation, whereas *fundamental resea* is mainly concerned with generalisations and with the formulation of a theory. Gather knowledge for knowledge's sake is termed fundamental research. Research concern some natural phenomenon or relating to pure mathematics are examples of fundame research. Similarly, research studies, concerning human behaviour carried on with a vi to make generalisations about human behaviour, are also examples of fundamental resea However, research aimed at certain conclusions facing a concrete social or business probl is an example of applied research. Research to identify social, economic or political tre that may affect a particular institution, marketing research, evaluation research are examp of applied research. Thus, the central aim of applied research is to discover a solution some pressing practical problems, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.

(iii) *Quantitative vs. Qualitative:* Quantitative research is based on the quantitative measurements of some characterstics. It is applicable to phenomena that can be expressed in terms of quantities. Qualitative research, on the other hand, is concerned with qualitative phenomenon, i.e., phenomena relating to or involving quality or kind. For instance, when we are interested in investigating the reasons for human behaviour (i.e., why people think or do certain things), we quite often talk of 'Motivation Research', an important type of qualitative research. This type of research aims at discovering the underlying motives and desires, using in depth interviews for the purpose. Other techniques of such research are word association tests, sentence completion tests, story completion tests and similar other projective techniques. Attitude or opinion research i.e., research designed to find out how people feel or what they think about a particular subject or institution is also qualitative research. Qualitative research is specially important in the behavioural sciences where the aim is to discover the underlying motives of human behaviour. Through such research we can analyse the various factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing. It may be stated, however, that to apply for qualitative research in practice is relatively a difficult job and therefore, while doing such research, one should seek guidance from experimental psychologists.

(iv) *Conceptual vs. Empirical:* Conceptual research is that related to some abstract idea(s) or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones. On the other hand, empirical research relies on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. We can also call it as experimental type of research. In such a research it is necessary to get facts at firsthand, at their source, and actively to go about doing certain things to stimulate the production of desired information. In such a research, the researcher must first provide himself with a working hypothesis or guess as to the probable results. He then works to get enough facts (data) to prove or disprove his hypothesis. He then sets up experimental designs which he thinks will manipulate the persons or the materials concerned so as to bring forth the desired information. Such research is thus characterised by the experimenter`s control over the variables under study and his deliberate manipulation of one of them to study its effects. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies are considered to be the most powerful support possible for testing a given hypothesis.

(v) *Some Other Types of Research:* All other types of research are variations of one or more of the above stated approaches, based on either the purpose of research, or the time required to accomplish research, on the environment in which research is done, or on the basis of some other similar factors. Form the point of view of time, we can think of research either as *one-time research or longitudinal research.* In the former case the research is confined to a single time-period, whereas in the latter case the research is carried on over several time-periods. Research can be *field-setting research or laboratory research* *simulation research*, depending upon the environment in which it is to be carried Research can as well be understood as *clinical or diagnostic research*. Such resea follow case-study methods or indepth approaches to reach the basic casual relations. Su studies usually go deep into the causes of things or events that interest us, using very sm samples and very deep probing data gathering devices. The research may be *explorat* or it may be formalized. The objective of exploratory research is the development hypotheses rather than their testing, whereas formalized research studies are those w substantial structure and with specific hypotheses to be tested. *Historical research* is t which utilizes historical sources like documents, remains, etc. to study events or idea the past, including the philosophy of persons and groups at any remote point of time. Resea can also be classified as *conclusion-oriented* and *decision-oriented*. While do conclusion-oriented research, a researcher is free to pick up a problem, redesign the enqu as he proceeds and is prepared to conceptualize as he wishes. Decision-oriented resea is always for the need of a decision maker and the researcher in this case is not fre embark upon research according to his own inclination. Operations research is an exam of decision oriented research since it is a scientific method of providing executive departm with a quantitative basis for decisions regarding operations under their control.

1.4 RESEARCH APPROACHES

The above description of the types of research brings to light the fact that there are two b approaches to research, viz., *quantitative approach* and the *qualitative approach*. The for involves the generation of data in quantitative form which can be subjected to rigorous quantita analysis in a formal and rigid fashion. This approach can be further sub-classified into *inferent* *experimental* and *simulation approaches* to research. The purpose of *inferential approach* is form a data base to infer characteristics or relationships of population. This usually means sur research where a sample of population is studied (questioned or observed) to determine characteristics, and it is then inferred that the population has the same characteristics. *Experimen* *approach* is characterised by much greater control over the research environment and in this c some variables are manipulated to observe their effect on other variables. *Simulation appro* involves the construction of an artificial environment within which relevant information and data be generated. This permits an observation of the dynamic behaviour of a system (or its sub-syst under controlled conditions. The term 'simulation' in the context of business and social scien applications refers to "the operation of a numerical model that represents the structure of a dyna process. Given the values of initial conditions, parameters and exogenous variables, a simulatio run to represent the behaviour of the process over time." Simulation approach can also be usef building models for understanding future conditions.

Qualitative approach to research is concerned with subjective assessment of attitudes, opini and behaviour. Research in such a situation is a function of researcher's insights and impressi Such an approach to research generates results either in non-quantitative form or in the form w is not subjected to rigorous quantitative and analysis. Generally, the techniques of focus group intervi projective techniques and depth interviews are used. All these are explained at length in chapters follow.

1.5 SIGNIFICANCE OF RESEARCH

All progress is born of inquiry. Doubt is often better than overconfidence, for it leads to inquiry, and quiry leads to invention” is a famous Hudson Maxim in context of which the significance of research an well be understood. Increased amounts of research make progress possible. *Research inculcates cientific and inductive thinking and it promotes the development of logical habits of thinking nd organisation.*

The role of research in several fields of applied economics, whether related to business or the economy as a whole, has greatly increased in modern times. The increasingly complex ature of business and governance has focussed attention on the use of research in solving operational roblems. Research, as an aid to economic policy, has gained added importance, both for governance nd business.

Research provides the basis for nearly all government policies in our economic system. or instance, government’s budgets rest in part on an analysis of the needs and desires of the people nd on the availability of revenues to meet those needs. The cost of needs has to be equated to robable revenues and this is a field where research is most needed. Through research we can levise alternative policies and can as well examine the consequences of each of these alternatives. ecision-making may not be a part of research, but research certainly facilitates the decisions of the olicy maker. Government has to chalk out programmes for dealing with all facets of the country’s arious operations and most of these are related directly or indirectly to economic conditions. The light of cultivators, the problems of big and small business and industry, working conditions, trade nion activities, the problems of distribution, even the size and nature of defence services are matters equiring research. Thus, research is considered necessary with regard to the allocation of nation’s esources. Another area in government, where research is necessary, is collecting information on the economic and social structure of the nation. Such information indicates what is happening in the economy and what changes are taking place. Collecting such statistical information is by no means a outine task, but it involves a variety of research problems. These days nearly all governments aintain large staff of research technicians or experts to carry on this work. Thus, in the context of government, research as a tool to economic policy has three distinct phases of operation, viz., i) investigation of economic structure through continual compilation of facts; (ii) diagnosis of events hat are taking place and the analysis of the forces underlying them; and (iii) the prognosis, i.e., the prediction of future developments.

Research has its special significance in solving various operational and planning problems of business and industry. Operations research and market research, along with motivational research, are considered crucial and their results assist, in more than one way, in taking business decisions. Market research is the investigation of the structure and development of a market for the purpose of formulating efficient policies for purchasing, production and sales. Operations research refers to the application of mathematical, logical and analytical techniques to the solution of business problems of cost minimisation or of profit maximisation or what can be termed as optimisation problems. Motivational research of determining why people behave as they do is mainly concerned with market characteristics. In other words, it is concerned with the determination of motivations underlying the consumer (market) behaviour. All these are of great help to people in business and industry who are responsible for taking business decisions. Research with regard to demand and market factors has great utility in business. Given knowledge of future demand, it is generally not difficult for a firm, or for an industry to adjust its supply schedule within the limits of its projected capacity. Market analysis has beco an integral tool of business policy these days. Business budgeting, which ultimately results ir projected profit and loss account, is based mainly on sales estimates which in turn depends business research. Once sales forecasting is done, efficient production and investment programm can be set up around which are grouped as the purchasing and financing plans. Research, th replaces intuitive business decisions by more logical and scientific decisions.

Research is equally important for social scientists in studying social relationships and seeking answers to various social problems. It provides the intellectual satisfaction of knowin few things just for the sake of knowledge and also has practical utility for the social scientist to kn for the sake of being able to do something better or in a more efficient manner. Research in soc sciences is concerned with (i) the development of a body of principles that helps in understandi the whole range of human interactions, and (ii) the practical guidance in solving immediate proble of human relations.

In addition to what has been stated above, the significance of research can also be understo keeping in view the following points:

(a) To those students who are to write a master's or Ph.D. thesis, research may mean careerism or a way to attain a high position in the social structure;
(b) To professionals in research methodology, research may mean a source of livelihood;
(c) To philosophers and thinkers, research may mean the outlet for new ideas and insights;
(d) To literary men and women, research may mean the development of new styles and creati work; and
(e) To analysts and intellectuals, research may mean the development of new theories.

Thus, research is the fountain of knowledge for the sake of knowledge and an important sour of providing guidelines for solving different business, governmental and social problems. It is a sort formal training which enables one to understand the new developments in one's field in a better wa

1.6 RESEARCH METHODS VERSUS METHODOLOGY

It seems appropriate at this juncture to explain the difference between research methods and resear methodology. *Research methods* may be understood as all those methods/techniques that are us for conduction of research. *Research methods or techniques, thus, refer to the methods the researche use in performing research operations*. In other words, all those methods which are used by t researcher during the course of studying his research problem are termed as research metho Since the object of research, particularly the applied research, it to arrive at a solution for a giv problem, the available data and the unknown aspects of the problem have to be related to each oth to make a solution possible. Keeping this in view, research methods can be put into the followi three groups:

1. In the first group we include those methods which are concerned with the collection data. These methods will be used where the data already available is not sufficient to arri at the required solution;

2. The second group consists of those statistical techniques which are used for establishing relationships between the data and the unknowns;
3. The third group consists of those methods which are used to evaluate the accuracy of the results obtained.

Research methods falling in the above stated last two groups are generally taken as the analytical ools of research.

At times, a distinction is also made between research techniques and research methods. *Research echniques* refer to the behaviour and instruments we use in performing research operations such as making observations, recording data, techniques of processing data and the like. *Research methods* refer to the behaviour and instruments used in selecting and constructing research technique. For instance, the difference between methods and techniques of data collection can better be understood from the details given in the following chart:

Type	*Methods*	*Techniques*
1. Library Research	(i) Analysis of historical records	Recording of notes, Content analysis, Tape and Film listening and analysis.
	(ii) Analysis of documents	Statistical compilations and manipulations, reference and abstract guides, contents analysis.
2. Field Research	(i) Non-participant direct observation	Observational behavioural scales, use of score cards, etc.
	(ii) Participant observation	Interactional recording, possible use of tape recorders, photographic techniques.
	(iii) Mass observation	Recording mass behaviour, interview using independent observers in public places.
	(iv) Mail questionnaire	Identification of social and economic background of respondents.
	(v) Opinionnaire	Use of attitude scales, projective techniques, use of sociometric scales.
	(vi) Personal interview	Interviewer uses a detailed schedule with open and closed questions.
	(vii) Focussed interview	Interviewer focuses attention upon a given experience and its effects.
	(viii) Group interview	Small groups of respondents are interviewed simultaneously.
	(ix) Telephone survey	Used as a survey technique for information and for discerning opinion; may also be used as a follow up of questionnaire.
	(x) Case study and life history	Cross-sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.
3. Laboratory Research	Small group study of random behaviour, play and role analysis	Use of audio-visual recording devices, use of observers, etc.

From what has been stated above, we can say that methods are more general. It is the method that generate techniques. However, in practice, the two terms are taken as interchangeable and when we talk of research methods we do, by implication, include research techniques within their compass.

Research methodology is a way to systematically solve the research problem. It may be understood as a science of studying how research is done scientifically. In it we study the various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them. It is necessary for the researcher to know not only the research methods/techniques but also the methodology. Researchers not only need to know how to develop certain indices or tests, how to calculate the mean, the mode, the median or the standard deviation or chi-square, how to apply particular research techniques, but they also need to know which of these methods or techniques, are relevant and which are not, and what would they mean and indicate. Researchers also need to understand the assumptions underlying various techniques and they need to know the criteria which they can decide that certain techniques and procedures will be applicable to certain proble and others will not. All this means that it is necessary for the researcher to design a methodology f his problem as the same may differ from problem to problem. For example, an architect, who desig a building, has to consciously evaluate the basis of his decisions, i.e., he has to evaluate why and what basis he selects particular size, number and location of doors, windows and ventilators, us particular materials and not others and the like. Similarly, in research the scientist has to expose t research decisions to evaluation before they are implemented. He has to specify very clearly a precisely what decisions he selects and why he selects them so that they can be evaluated by others als

From what has been stated above, we can say that research methodology has many dimensio and research methods do constitute a part of the research methodology. The scope of resear methodology is wider than that of research methods. *Thus, when we talk of research methodolo we not only talk of the research methods but also consider the logic behind the methods we u in the context of our research study and explain why we are using a particular method technique and why we are not using others so that research results are capable of bei evaluated either by the researcher himself or by others*. Why a research study has been undertake how the research problem has been defined, in what way and why the hypothesis has been formulate what data have been collected and what particular method has been adopted, why particular techniq of analysing data has been used and a host of similar other questions are usually answered when talk about research methodology concerning a research problem or study.

1.7 RESEARCH AND SCIENTIFIC METHOD

For a clear perception of the term research, one should know the meaning of scientific method. T two terms, research and scientific method, are closely related. In research, we study the natu reasons, and the consequences of a set of circumstances which are controlled experimentally observed as they appear. Researcher is usually interested in particular results, the repetitions those results, and generalizations. On the other hand, the philosophy common to all research metho and techniques, although they may vary considerably from one science to another, is usually giv the name of scientific method. In this context, Karl Pearson writes, “The scientific method is one a same in the branches (of science) and that method is the method of all logically trained minds … t unity of all sciences consists alone in its methods, not its material; the man who classifies facts of a kind whatever, who sees their mutual relation and describes their sequences, is applying the scienti method and is a man of science.” Scientific method is the pursuit of truth as determined by logic considerations. The ideal of science is to achieve a systematic interrelation of facts. Scientific meth attempts to achieve this ideal by experimentation, observation, logical arguments from accept postulates and a combination of these three in varying proportions. In scientific method, logic aids formulating propositions explicitly and accurately so that their possible alternatives become cle Further, logic develops the consequences of such alternatives, and when these are compared w observable phenomena, it becomes possible for the researcher or the scientist to state which alternati is the most in harmony with the observed facts. All this is done through experimentation and surv investigations which constitute the integral parts of scientific method.

Experimentation is done to test hypotheses and to discover new relationships, if any, among ariables. However, sometimes the conclusions drawn on the basis of experimental data may be isleading for either faulty assumptions, poorly designed experiments, badly executed experiments r faulty interpretations. As such the researcher must pay all possible attention while developing the xperimental design and drawing inferences. The purpose of survey investigations may also be to rovide scientifically gathered information to work as a basis for the researchers for their conclusions.

The scientific method is, thus, based on certain basic postulates which can be stated as under:

1. It relies on empirical evidence;
2. It utilizes relevant concepts;
3. It is committed to only objective considerations;
4. It aims at nothing but making only adequate and correct statements about population objects;
5. It results into probabilistic predictions;
6. Its methodology is made known to all concerned for critical scrutiny and are for use in testing the conclusions through replication;
7. It aims at formulating most general axioms or what can be termed as scientific theories.

Thus, the scientific method encourages a rigorous, method wherein the researcher is guided by he rules of logical reasoning, a method wherein the investigation proceeds in an orderly manner and method that implies internal consistency.

1.8 RESEARCH PROCESS

Before embarking on the details of research methodology and techniques, it seems appropriate to present a brief overview of the research process. Research process consists of series of actions or steps necessary to effectively carry out research. The chart shown in Figure 1.1 illustrates a research process.

The chart indicates that the research process consists of a number of closely related activities, as shown through I to VII. But such activities overlap continuously rather than following a strictly prescribed sequence. At times, the first step determines the nature of the last step to be undertaken. If subsequent procedures have not been taken into account in the early stages, serious difficulties may arise which may even prevent the completion of the study. One should remember that neither various steps involved in a research process are mutually exclusive; nor they are separate and distinct. They do not necessarily follow each other in any specific order and the researcher has to be constantly anticipating at each step in the research process the requirements of the subsequent steps. However, the following order concerning various steps provides a useful procedural guideline regarding the research process: (i) formulating the research problem; (ii) extensive literature survey; (iii) developing the hypothesis; (iv) preparing the research design; (v) determining sample design; (vi) collecting the data; (vii) execution of the project; (viii) analysis of data; (ix) hypothesis testing; (x) generalisations and interpretation; and (xi) preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached.

A brief description of the above stated steps will be helpful.

RESEARCH PROCESS IN FLOW CHART

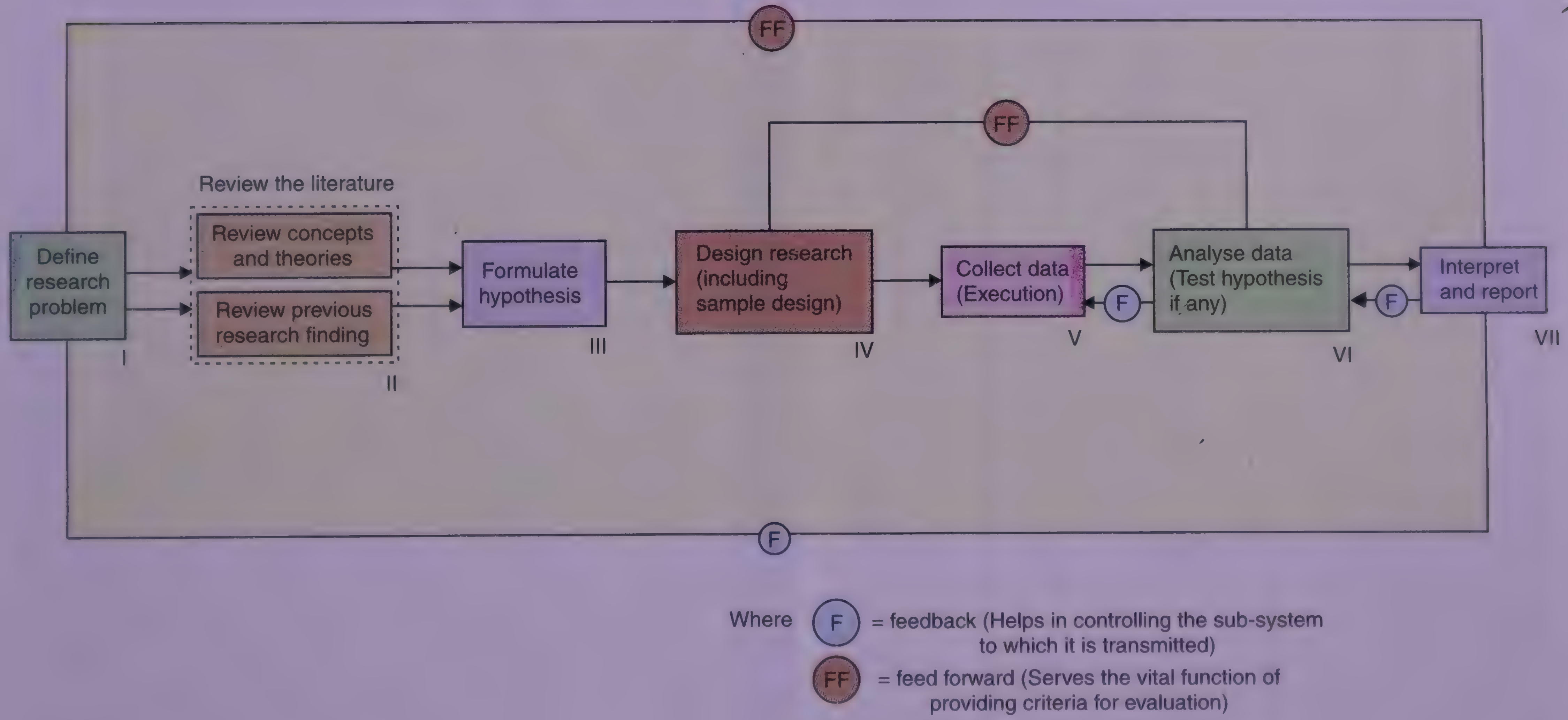

FF
Review the literature
Define research problem
I
Review concepts and theories
Review previous research finding
II
Formulate hypothesis
III
Design research (including sample design)
IV
FF
Collect data (Execution)
V
F
Analyse data (Test hypothesis if any)
VI
F
Interpret and report
VII
F
Where
F = feedback (Helps in controlling the sub-system to which it is transmitted)
FF = feed forward (Serves the vital function of providing criteria for evaluation)

Fig. 1.1

(i) **Formulating the research problem:** There are two types of research problems, viz., those which relate to states of nature and those which relate to relationships between variables. At the very outset, the researcher must single out the problem he wants to study, i.e., he must decide the general area of interest or aspect of a subject-matter that he would like to inquire into. Initially, the problem may be stated in a broad general way and then the ambiguities, if any, relating to the problem be resolved. Then, the feasibility of a particular solution has to be considered before a working formulation of the problem can be set up. The formulation of a general topic into a specific research problem, thus, constitutes the first step in a scientific enquiry. Essentially two steps are involved in formulating the research problem, viz., understanding the problem thoroughly, and rephrasing the same into meaningful terms from an analytical point of view.

The best way of understanding the problem is to discuss it with one's own colleagues or with those having some expertise in the matter. Often, a problem is put in general terms and it is up to the researcher to narrow it down and phrase the problem in operational terms. In private business units or in governmental organisations, the problem is usually earmarked by the administrative agencies with whom the researcher can discuss as to how the problem originally came out and what considerations are involved in its possible solutions.

The researcher must at the same time examine all available literature to get himself acquainted with the selected problem. He may review two types of literature—the conceptual literature concerning the concepts and theories, and the empirical literature consisting of studies made earlier which are similar to the one proposed. The basic outcome of this review will be the knowledge as to what data and other materials are available for operational purposes which will enable the researcher to specify his own research problem in a meaningful context. After this the researcher rephrases the problem into analytical or operational terms i.e., to put the problem in as specific terms as possible. This task of formulating, or defining a research problem is a step of greatest importance in the entire research process. The problem to be investigated must be defined unambiguously for that will help discriminating relevant data from irrelevant ones. Care must, however, be taken to verify the objectivity and validity of the background facts concerning the problem. The statement of the objective is of basic importance because it determines the data which is to be collected, the characteristics of the data which is relevant, relations which are to be explored, the choice of techniques to be used in these explorations and the form of the final report. If there are certain pertinent terms, the same should be clearly defined along with the task of formulating the problem. In fact, formulation of the problem often follows a sequential pattern where a number of formulations are set up, each formulation more specific than the preceeding one, each one phrased in more analytical terms, and each more realistic in terms of the available data and resources.

(ii) **Extensive literature survey:** Once the problem is formulated, a brief summary of it should be written down. At this juncture the researcher should undertake extensive literature survey connected with the problem. For this purpose, the abstracting and indexing journals and published or unpublished bibliographies are the first place to go to. Academic journals, conference proceedings, government reports, books etc., must be tapped depending on the nature of the problem. In this process, it should be remembered that one source will lead to another. The earlier studies, if any, which are similar to the study in hand should be carefull studied. A good library will be of a great help to the researcher at this stage. In this era c internet, it has become quite easy to search the articles. One can use some search engin like 'google' or specially 'google scholars' to search scholarly articles.

(iii) **Development of working hypotheses:** After extensive literature survey, research should state in clear terms the working hypothesis or hypotheses. Working hypothesis i tentative assumption made in order to draw out and test its logical or empirical consequence As such the manner in which research hypotheses are developed is particularly importan since they provide the focal point for research. They also affect the manner in which test must be conducted in the analysis of data and indirectly the quality of data which is require for the analysis. Hypothesis should be very specific and limited to the piece of research i hand because it has to be tested. The role of the hypothesis is to guide the researcher b delimiting the area of research and to keep him on the right track. It sharpens his thinkin and focuses attention on the more important facets of the problem. It also indicates th type of data required and the type of methods of data analysis to be used.

How does one go about developing working hypotheses? The answer is by using the followin approach:

(a) Discussions with colleagues and experts about the problem, its origin and the objectiv in seeking a solution;

(b) Examination of data and records, if available, concerning the problem for possib trends, peculiarities and other clues;

(c) Review of similar studies in the area or of the studies on similar problems; and

(d) Exploratory personal investigation which involves original field interviews on a limite scale with interested parties and individuals with a view to secure greater insight in the practical aspects of the problem.

Thus, working hypotheses arise as a result of a priori thinking about the subject, examinatio of the available data and material including related studies and the counsel of experts an interested parties. Working hypotheses are more useful when stated in precise and clear defined terms. It may as well be remembered that occasionally we may encounter a proble where we do not need working hypotheses, specially in the case of exploratory or formulati researches which do not aim at testing the hypothesis. But as a general rule, specificatio of working hypotheses is another basic step of the research process is many researc problems.

(iv) **Preparing the research design:** The research problem having been formulated in cle terms, the researcher will be required to prepare a research design, i.e., he will have state the conceptual structure within which research would be conducted. The preparatio of such a design facilitates research to be as efficient as possible yielding maxim information. In other words, the function of research design is to provide for the collectio of relevant evidence with optimum effort, time and expenditure. But how all these can achieved depends mainly on the research purpose. Research purposes may be group into four categories, viz., (i) Exploration, (ii) Description, (iii) Diagnosis, a (iv) Experimentation. A flexible research design which provides opportunity for consideri many different aspects of a problem is considered appropriate if the purpose of the research study is that of exploration. But when the purpose happens to be an accurate description of a situation or of an association between variables, the suitable design will be one that minimises bias and maximises the reliability of the data collected and analysed.

There are several research designs, such as, experimental and non-experimental hypothesis testing. Experimental designs can be either informal designs (such as before-and-after without control, after-only with control, before-and-after with control) or formal designs (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial designs), out of which the researcher must select one for his own project.

The preparation of the research design, appropriate for a particular research problem, involves usually the consideration of the following:

(a) the means of obtaining the information;
(b) the availability and skills of the researcher and his staff (if any);
(c) explanation of the way in which selected means of obtaining information will be organised and the reasoning leading to the selection;
(d) the time available for research; and
(e) the cost factor relating to research, i.e., the finance available for the purpose.

(v) **Determining sample design:** All the items under consideration in any field of inquiry constitute a 'universe' or 'population'. A complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry when all the items are covered, no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of observations increases. Moreover, there is no way of checking the element of bias or its extent except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Not only this, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sample.

The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In other words, a sample design is a definite plan determined before any data is actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city's 200 drugstores in a certain way constitutes a sample design. Samples can be either probability samples or non-probability samples. With probability samples each element has a known probability of being included in the sample but the non-probability samples do not allow the researcher to determine this probability. Probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples are those based on convenience sampling, judgement sampling and quotá sampling techniques. A brief mention of the important sample designs is as follows:

(a) *Deliberate sampling:* Deliberate sampling is also known as purposive or non-probabili sampling. This sampling method involves purposive or deliberate selection of particul units of the universe for constituting a sample which represents the universe. When populati elements are selected for inclusion in the sample based on the ease of access, it can called *convenience sampling*. If a researcher wishes to secure data from, say, gasoli buyers, he may select a fixed number of petrol stations and may conduct interviews these stations. This would be an example of convenience sample of gasoline buyers. times such a procedure may give very biased results particularly when the population is n homogeneous. On the other hand, in *judgement sampling* the researcher's judgement used for selecting items which he considers as representative of the population. For examp a judgement sample of college students might be taken to secure reactions to a new meth of teaching. Judgement sampling is used quite frequently in qualitative research where t desire happens to develop hypothesis rather than to generalise to larger populations.

(b) *Simple random sampling:* This type of sampling is also known as chance sampling probability sampling where each and every item in the population has an equal chance inclusion in the sample and each one of the possible samples, in case of finite universe, H the same probability of being selected. For example, if we have to select a sample of 3 items from a universe of 15,000 items, then we can put the names or numbers of all t 15,000 items on slips of paper and conduct a lottery. Using the random number tables another method of random sampling. To select the sample, each item is assigned a numb from 1 to 15,000. Then, 300 five digit random numbers are selected from the table. To this we select some random starting point and then a systematic pattern is used in proceedi through the table. We might start in the 4th row, second column and proceed down t column to the bottom of the table and then move to the top of the next column to the rig When a number exceeds the limit of the numbers in the frame, in our case over 15,000, i simply passed over and the next number selected that does fall within the relevant ran Since the numbers were placed in the table in a completely random fashion, the resulti sample is random. This procedure gives each item an equal probability of being selected. case of infinite population, the selection of each item in a random sample is controlled the same probability and that successive selections are independent of one another.

(c) *Systematic sampling:* In some instances the most practical way of sampling is to sel every 15th name on a list, every 10th house on one side of a street and so on. Sampling this type is known as systematic sampling. An element of randomness is usually introduc into this kind of sampling by using random numbers to pick up the unit with which to sta This procedure is useful when sampling frame is available in the form of a list. In suc design the selection process starts by picking some random points in the list and then ev *n*th element is selected until the desired number is secured.

(d) *Stratified sampling:* If the population from which a sample is to be drawn does not constit a homogeneous group, then stratified sampling technique is applied so as to obtai representative sample. In this technique, the population is stratified into a number of n overlapping subpopulations or strata and sample items are selected from each stratum the items selected from each stratum is based on simple random sampling the entire procedu first stratification and then simple random sampling, is known as *stratified random sampli*

(e) *Quota sampling:* In stratified sampling the cost of taking random samples from individual strata is often so expensive that interviewers are simply given quota to be filled from different strata, the actual selection of items for sample being left to the interviewer's judgement. This is called *quota sampling*. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population. Quota sampling is thus an important form of non-probability sampling. Quota samples generally happen to be judgement samples rather than random samples.

(f) *Cluster sampling and area sampling:* Cluster sampling involves grouping the population and then selecting the groups or the clusters rather than individual elements for inclusion in the sample. Suppose some departmental store wishes to sample its credit card holders. It has issued its cards to 15,000 customers. The sample size is to be kept say 450. For cluster sampling this list of 15,000 card holders could be formed into 100 clusters of 150 card holders each. Three clusters might then be selected for the sample randomly. The sample size must often be larger than the simple random sample to ensure the same level of accuracy because cluster sampling procedural potential for order bias and other sources of error is usually accentuated. The clustering approach can, however, make the sampling procedure relatively easier and increase the efficiency of field work, specially in the case of personal interviews.

Area sampling is quite close to cluster sampling and is often talked about when the total geographical area of interest happens to be big one. Under area sampling we first divide the total area into a number of smaller non-overlapping areas, generally called geographical clusters, then a number of these smaller areas are randomly selected, and all units in these small areas are included in the sample. Area sampling is specially helpful where we do not have the list of the population concerned. It also makes the field interviewing more efficient since interviewer can do many interviews at each location.

(g) *Multi-stage sampling:* This is a further development of the idea of cluster sampling. This technique is meant for big inquiries extending to a considerably large geographical area like an entire country. Under multi-stage sampling the first stage may be to select large primary sampling units such as states, then districts, then towns and finally certain families within towns. If the technique of random-sampling is applied at all stages, the sampling procedure is described as multi-stage random sampling.

(h) *Sequential sampling:* This is somewhat a complex sample design where the ultimate size of the sample is not fixed in advance but is determined according to mathematical decisions on the basis of information yielded as survey progresses. This design is usually adopted under acceptance sampling plan in the context of statistical quality control.

In practice, several of the methods of sampling described above may well be used in the same study in which case it can be called mixed sampling. It may be pointed out here that normally one should resort to random sampling so that bias can be eliminated and sampling error can be estimated. But purposive sampling is considered desirable when the universe happens to be small and a known characteristic of it is to be studied intensively. Also, there are conditions under which sample designs other than random sampling may be considered better for reasons like convenience and low costs. *The sample design to be used must be decided by the researcher taking into consideration the nature of the inquiry and other related factors.*

(vi) **Collecting the data:** In dealing with any real life problem it is often found that data a hand is inadequate, and hence, it becomes necessary to collect data that is appropriate There are several ways of collecting the appropriate data which differ considerably in context of cost, time and other resources at the disposal of the researcher.

Primary data can be collected either through experiment or through survey. If the researcher conducts an experiment, he observes some quantitative measurements, or the data, with the help of which he examines the truth contained in his hypothesis. But in the case of a survey, data can be collected by any one or more of the following ways:

(i) *By observation:* This method implies the collection of information by way of investigator's own observation, without interviewing the respondents. The information obtained relates to what is currently happening and is not complicated by either the past behaviour or future intentions or attitudes of respondents. This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is not suitable in inquiries where large samples are concerned.

(ii) *Through personal interview:* The investigator follows a rigid procedure and seeks answers to a set of pre-conceived questions through personal interviews. This method of collecting data is usually carried out in a structured way where output depends upon the ability of the interviewer to a large extent.

(iii) *Through telephone interviews:* This method of collecting information involves contacting the respondents on telephone itself. This is not a very widely used method but it plays an important role in industrial surveys in developed regions, particularly, when the survey has to be accomplished in a very limited time.

(iv) *By mailing of questionnaires:* The researcher and the respondents do come in contact with each other if this method of survey is adopted. Questionnaires are mailed to the respondents with a request to return after completing the same. It is the most extensively used method in various economic and business surveys. Before applying this method, usually a Pilot Study for testing the questionnaire is conducted which reveals the weaknesses, if any, of the questionnaire. Questionnaire to be used must be prepared very carefully so that it may prove to be effective in collecting the relevant information.

(v) *Through schedules:* Under this method the enumerators are appointed and given training. They are provided with schedules containing relevant questions. These enumerators go to respondents with these schedules. Data is collected by filling up the schedules by enumerators on the basis of replies given by respondents. Much depends upon the capability of enumerators so far as this method is concerned. Some occasional field checks on the work of the enumerators may ensure sincere work.

The researcher should select one of these methods of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy. Though he should pay attention to all these factors but much depends upon the ability and experience of the researcher. In this context Dr. A.L. Bowley very aptly remarks that in collection of statistical data commonsense is the chie requisite and experience of the chief teacher.

(vii) Execution of the project: Execution of the project is a very important step in the research process. If the execution of the project proceeds on correct lines, the data to be collected would be adequate and dependable. The researcher should see that the project is executed in a systematic manner and or time. If the survey is to be conducted by means of structured questionnaires, data can be readily machine processed. In such a situation, questions as well as the possible answers may be coded. If the data is to be collected through interviewers, arrangements should be made for proper selection and training of the interviewers. The training may be given with the help of instruction manuals which explain clearly the job of the interviewers at each step. Occasional field checks should be made to ensure that the interviewers are doing their assigned job sincerely and efficiently. A careful watch should be kept for unanticipated factors in order to keep the survey realistic as much as possible. This, in other words, means that steps should be taken to ensure that the survey is under statistical control so that the collected information is in accordance with the pre-defined standard of accuracy. If some of the respondents do not cooperate, some suitable methods should be designed to tackle this problem. One method of dealing with the non-response problem is to make a list of the non-respondents and take a small sub-sample of them, and then with the help of experts vigorous efforts can be made for securing response.

(viii) Analysis of data: After the data has been collected, the researcher turns to the task of analysing them. The analysis of data requires a number of closely related operations such as establishment of categories, the application of these categories to raw data through coding, tabulation and then drawing statistical inferences. The unwieldy data should necessarily be condensed into a few manageable groups and tables for further analysis. Thus, researcher should classify the raw data into some purposeful and usable categories. *Coding* operation is usually done at this stage through which the categories of data is transformed into symbols that may be tabulated and counted. *Editing* is the procedure that improves the quality of the data for coding. With coding the stage is ready for tabulation. *Tabulation* is a part of the technical procedure wherein the classified data is put in the form of tables. The mechanical devices can be made use of at this juncture. A great deal of data, specially in large inquiries, is tabulated by computers. Computers not only save time but also make it possible to study large number of variables affecting a problem simultaneously.

Analysis work after tabulation is generally based on the computation of various percentages, coefficients, etc., by applying various well defined statistical formulae. In the process of analysis, relationships or differences supporting or conflicting with original or new hypothesis should be subjected to tests of significance to determine with what validity data can be said to indicate any conclusion(s). For instance, if there are two samples of weekly wages, each sample being drawn from factories in different parts of the same city, giving two different mean values, then our problem may be whether the two mean values are significantly different or the difference is just a matter of chance. Through the use of statistical tests we can establish whether such a difference is a real one or is the result of random fluctuations. If the difference happens to be real, the inference will be that the two samples come from different universe and if the difference is due to chance, the conclusion would be that the two samples belong to the same universe. Similarly, the technique of analysis of variance can help us in analysing whether three or more varieties of seeds grown on certain fields yield significantly different results or not. In brief, the researcher can analyse the collected data with the help of various statistical measures.

(ix) **Hypothesis-testing:** After analysing the data as stated above, the researcher is in position to test the hypothesis, if any, he had formulated earlier. Do the facts support th hypothesis or they happen to be contrary? This is the usual question which should b answered while testing hypothesis. Various tests, such as Chi square test, t-test, F-test have been developed by statisticians for the purpose. These tests are discussed in detai later in the book. The hypothesis may be tested through the use of one or more of suc tests, depending upon the nature and object of research inquiry. Hypothesis-testing will result in either accepting the hypothesis or in rejecting it. If the researcher had no hypothesi to start with, generalisations established on the basis of data may be stated as hypothesis t be tested by subsequent researches in times to come.

(x) **Generalisations and interpretation:** If a hypothesis is tested and upheld several times it may be possible for the researcher to arrive at generalisation, i.e., to build a theory. As matter of fact, the real value of research lies in its ability to arrive at certain generalisations If the researcher had no hypothesis to start with, he might seek to explain his findings o the basis of some theory. It is known as interpretation. The process of interpretation ma quite often trigger off new questions which in turn may lead to further researches.

(xi) **Preparation of the report or the thesis:** Finally, the researcher has to prepare th report of what has been done by him. Writing of report must be done with great car keeping in view the followings:

(a) The layout of the report should be as follows: the preliminary pages; the main text, an the end matter.

In its preliminary pages the report should carry title and date followed by acknowledgement and foreword. Then there should be a table of contents followed by a list of tables and list o illustrations, if any, given in the report.

The main text of the report should have the following parts:

- *Introduction:* It should contain a clear statement of the objective of the research and a explanation of the methodology adopted in accomplishing the research. The scope of th study along with various limitations should as well be stated in this part.
- *Summary of findings:* After introduction there would appear a statement of findings an recommendations in non-technical language. If the findings are extensive, they should b summarised.
- *Main report:* The main body of the report should be presented in logical sequence an broken-down into readily identifiable sections.
- *Conclusion:* Towards the end of the main text, researcher should again put down th results of his research clearly and precisely. In fact, it is the final summing up.

At the end of the report, appendices should be enlisted in respect of all technical data Bibliography, i.e., list of books, journals, reports, etc., consulted, should also be given at the end. Inde should also be given specially in a published research report.

(b) Report should be written in a concise and objective style in simple language avoiding vagu expressions such as 'it seems,' 'there may be', and the like.

(c) Charts and illustrations in the main report should be used only if they present the information more clearly and forcibly.

(d) Calculated 'confidence limits' must be mentioned and the various constraints experienced in conducting research operations may as well be stated.

1.9 CRITERIA OF GOOD RESEARCH

One expects scientific research to satisfy the following criteria:

(a) The purpose of the research should be clearly defined and common concepts be used.

(b) The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.

(c) The procedural design of the research should be carefully planned to yield results that are as objective as possible.

(d) The researcher should report with complete frankness, flaws in procedural design and estimate their effects upon the findings.

(e) The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.

(f) Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis.

(g) Greater confidence in research is warranted if the researcher is experienced, has a good reputation in research and is a person of integrity.

In other words, we can state the qualities of a good research as under:

(a) *Good research is systematic:* It means that research is structured with specified steps to be taken in a specified sequence in accordance with the well defined set of rules. Systematic characteristic of the research does not rule out creative thinking but it certainly does reject the use of guessing and intuition in arriving at conclusions.

(b) *Good research is logical:* This implies that research is guided by the rules of logical reasoning and the logical process of induction and deduction are of great value in carrying out research. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from some premise to a conclusion which follows from that very premise. In fact, logical reasoning makes research more meaningful in the context of decision making.

(c) *Good research is empirical:* It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.

(d) *Good research is replicable:* This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for decisions.

1.10 PROBLEMS ENCOUNTERED BY RESEARCHERS IN INDIA

Researchers in India, particularly those engaged in empirical research, are facing several problem
Some of the important problems are as follows:

(a) *The lack of a scientific training in the methodology of research* is a great impedimen for researchers in our country. There is a paucity of competent researchers. Man researchers take a leap in the dark without knowing research methods. Most of the work which goes in the name of research is not methodologically sound. Research to man researchers and even to their guides, is mostly a scissor and paste job without any insigh shed on the collated materials. The consequence is obvious, viz., the research results, quit often, do not reflect the reality or realities. Thus, a systematic study of research methodolog is an urgent necessity. Before undertaking research projects, researchers should be wel equipped with all the methodological aspects. As such, *efforts should be made to provid short-duration intensive courses for meeting this requirement.*

(b) There is *insufficient interaction* between the university research departments on one sid and business establishments, government departments and research institutions on the othe side. A great deal of primary data of non-confidential nature remains untouched/untreate by the researchers for want of proper contacts. *Efforts should be made to develop satisfactory liaison among all concerned for better and realistic researches.* There i need for developing some mechanisms of a university—industry interaction programme s that academics can get ideas from practitioners on what needs to be researched an practitioners can apply the research done by the academics.

(c) Most of the business units in our country do not have the confidence that the materia supplied by them to researchers will not be misused and as such they are often reluctant i supplying the needed information to researchers. The concept of secrecy seems to b sacrosanct to business organisations in the country so much that it proves an impermeable barrier to researchers. Thus, *there is the need for generating the confidence that th information/data obtained from a business unit will not be misused.*

(d) *Research studies overlapping one another are undertaken quite often for want o adequate information.* This results in duplication and fritters away resources. This problem can be solved by proper compilation and revision, at regular intervals, of a list of subjects on which and the places where the research is going on. Due attention should be given towards identification of research problems in various disciplines of applied science which are of immediate concern to the industries.

(e) *There does not exist a code of conduct for researchers* and inter-university and inter-departmental rivalries are also quite common. Hence, there is need for developing a code of conduct for researchers which, if adhered sincerely, can win over this problem.

(f) Many researchers in our country also face *the difficulty of adequate and timely secretarial assistance*, including computerial assistance. This causes unnecessary delays in the completion of research studies. All possible efforts to be made in this direction so that efficient secretarial assistance is made available to researchers and that too well in time. University Grants Commission must play a dynamic role in solving this difficulty.

(g) *Library management and functioning is not satisfactory at many places* and much of the time and energy of researchers is spent in tracing out the books, journals, reports, etc., rather than in tracing out relevant material from them.

(h) *There is also the problem that many of our libraries are not able to get copies of old and new Acts/Rules, reports and other government publications in time*. This problem is felt more in libraries which are away in places from Delhi and/or the state capitals. Thus, efforts should be made for the regular and speedy supply of all governmental publications to reach our libraries.

(i) *There is also the difficulty of timely availability of published data* from various government and other agencies doing this job in our country. Researcher also faces the problem on account of the fact that the published data vary quite significantly because of differences in coverage by the concerning agencies.

(j) There may, at times, take place *the problem of conceptualization* and also problems relating to the process of data collection and related things.

PROBLEMS

1. Briefly describe the different steps involved in a research process.
2. What do you mean by research? Explain its significance in modern times.
3. Distinguish between Research methods and Research methodology.
4. Describe the different types of research, clearly pointing out the difference between an experiment and a survey.
5. Write short notes on:
   (a) Design of the research project;
   (b) Ex post facto research;
   (c) Objectives of research;
   (d) Criteria of good research;
   (e) Research and scientific method.
6. "Empirical research in India in particular creates so many problems for the researchers". State the problems that are usually faced by such researchers.
7. "A research scholar has to work as a judge and derive the truth and not as a pleader who is only eager to prove his case in favour of his plaintiff." Discuss the statement pointing out the objectives of research.
8. "Creative management, whether in public administration or private industry, depends on methods of inquiry that maintain objectivity, clarity, accuracy and consistency". Discuss this statement and examine the significance of research".
9. "Research is much concerned with proper fact finding, analysis and evaluation." Do you agree with this statement? Give reasons in support of your answer.
10. It is often said that there is not a proper link between some of the activities under way in the world of academics and in most busineses in our country. Account for this state of affairs and give suggestions for improvement.

2

Chapter

Defining the Research Problem

In research process, the first and foremost step happens to be that of selecting and properly defining a research problem. A researcher must find the problem and formulate it so that it becomes susceptible to research. Like a medical doctor, a researcher must examine all the symptoms (presented to him or observed by him) concerning a problem before he can diagnose correctly. To define a problem correctly, a researcher must know: what a problem is?

2.1 WHAT IS A RESEARCH PROBLEM?

A research problem, in general, refers to some difficulty which a researcher experiences in the context of either a theoretical or practical situation and wants to obtain a solution for the same. Usually we say that a research problem does exist if the following conditions are met with:

(i) There must be an individual (or a group or an organisation), let us call it ‘*I*,’ to whom the problem can be attributed. The individual or the organisation, as the case may be, occupies an environment, say ‘*N*’, which is defined by values of the uncontrolled variables, Y_j.

(ii) There must be at least two courses of action, say C_1 and C_2, to be pursued. A course of action is defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

(iii) There must be at least two possible outcomes, say O_1 and O_2, of a course of action, of which one should be preferable to the other. In other words, this means that there must be at least one outcome that the researcher wants, i.e., an objective.

(iv) The courses of action available must provide some chance of obtaining the objective, but they cannot provide the same chance, otherwise the choice would not matter. Thus, if $P(O_i \mid I, C_j, N)$ represents the probability that an outcome O_i will occur, if *I* select C_j in *N*, then $P(O_1 \mid I, C_1, N) \neq P(O_1 \mid I, C_2, N)$. In simple words, we can say that the choices must have unequal efficiencies for the desired outcomes.

Over and above these conditions, the individual or the organisation can be said to have the problem only if ‘*I*’ does not know what course of action is best, i.e., ‘*I*’, must be in doubt about the solution. Thus, an individual or a group of persons can be said to have a problem which can be echnically described as a research problem, if they (individual or the group), having one or more esired outcomes, are confronted with two or more courses of action that have some but not equal fficiency for the desired objective(s) and are in doubt about which course of action is best.

We can, thus, state the components of a research problem as under:

(i) There must be an individual or a group which has some difficulty or the problem.
(ii) There must be some objective(s) to be attained at. If one wants nothing, one cannot have a problem.
(iii) There must be alternative means (or the courses of action) for obtaining the objective(s) one wishes to attain. This means that there must be *at least two means* available to a researcher for if he has no choice of means, he cannot have a problem.
(iv) There must remain some doubt in the mind of a researcher with regard to the selection of alternatives. This means that research must answer the question concerning the relative efficiency of the possible alternatives.
(v) There must be some environment(s) to which the difficulty pertains.

Thus, a research problem is one which requires a researcher to find out the best solution for the given problem, i.e., to find out by which course of action the objective can be attained optimally in the context of a given environment. There are several factors which may result in making the problem complicated. For instance, the environment may change affecting the efficiencies of the courses of action or the values of the outcomes; the number of alternative courses of action may be very large; persons not involved in making the decision may be affected by it and react to it favourably or unfavourably, and similar other factors. All such elements (or at least the important ones) may be thought of in context of a research problem.

2.2 SELECTING THE PROBLEM

The research problem undertaken for study must be carefully selected. The task is a difficult one, although it may not appear to be so. Help may be taken from a research guide in this connection. Nevertheless, every researcher must find out his own salvation for research problems cannot be borrowed. A problem must spring from the researcher's mind like a plant springing from its own seed. If our eyes need glasses, it is not the optician alone who decides about the number of the lens we require. We have to see ourselves and enable him to prescribe for us the right number by cooperating with him. Thus, a research guide can at the most only help a researcher choose a subject. However, the following points may be observed by a researcher in selecting a research problem or a subject for research:

(i) Subject which is overdone should not be normally chosen, for it will be a difficult task to throw any new light in such a case.
(ii) Controversial subject should not become the choice of an average researcher.
(iii) Too narrow or too vague problems should be avoided.
(iv) The subject selected for research should be familiar and feasible so that the related research material or sources of research are within one's reach. Even then it is quite difficult to supply definitive ideas concerning how a researcher should obtain ideas for his research. For this purpose, a researcher should contact an expert or a professor in the University

who is already engaged in research. He may as well read articles published in curren literature available on the subject and may think how the techniques and ideas discusse therein might be applied to the solution of other problems. He may discuss with others wh he has in mind concerning a problem. In this way he should make all possible efforts i selecting a problem.

(v) The importance of the subject, the qualifications and the training of a researcher, the cost involved, the time factor are few other criterias that must also be considered in selecting problem. In other words, before the final selection of a problem is done, a researcher mus ask himself the following questions:

(a) Whether he is well equipped in terms of his background to carry out the research?

(b) Whether the study falls within the budget he can afford?

(c) Whether the necessary cooperation can be obtained from those who must participat in research as subjects?

If the answers to all these questions are in the affirmative, one may become sure so far a the practicability of the study is concerned.

(vi) The selection of a problem must be preceded by a preliminary study. This may not be necessary when the problem requires the conduct of a research closely similar to one tha has already been done. But when the field of inquiry is relatively new and does not hav available a set of well developed techniques, a brief feasibility study must always be undertaken.

If the subject for research is selected properly by observing the above mentioned points, the research will not be a boring drudgery, rather it will be love's labour. In fact, zest for work is a must. The subject or the problem selected must involve the researcher and must have an upper most place in his mind so that he may undertake all pains needed for the study.

2.3 NECESSITY OF DEFINING THE PROBLEM

Quite often we all hear that a problem clearly stated is a problem half solved. This statement signifies the need for defining a research problem. The problem to be investigated must be defined unambiguously for that will help to discriminate relevant data from the irrelevant ones. A proper definition of research problem will enable the researcher to be on the track whereas an ill-defined problem may create hurdles. Questions like: What data is to be collected? What characteristics of data are relevant and need to be studied? What relations are to be explored. What techniques are to be used for the purpose? and similar other questions crop up in the mind of the researcher who can well plan his strategy and find answers to all such questions only when the research problem has been well defined. Thus, defining a research problem properly is a prerequisite for any study and is a step of the highest importance. In fact, formulation of a problem is often more essential than its solution. It is only on the careful detailing of the research problem that we can work out the research design and can smoothly carry on all the consequential steps involved while doing research.

2.4 TECHNIQUE INVOLVED IN DEFINING A PROBLEM

Let us start with the question: What does one mean when he/she wants to define a research problem? The answer may be that one wants to state the problem along with the bounds within which it is to be studied. In other words, defining a problem involves the task of laying down boundaries within which a researcher shall study the problem with a pre-determined objective in view.

How to define a research problem is undoubtedly a herculean task. However, it is a task that must be tackled intelligently to avoid the perplexity encountered in a research operation. The usual approach is that the researcher should himself pose a question (or in case someone else who wants the researcher to carry on research, the concerned individual, organisation or an authority should pose the question to the researcher) and set-up techniques and procedures for throwing light on the question concerned for formulating or defining the research problem. But such an approach generally does not produce definitive results because the question phrased in such a fashion is usually in broad general terms and as such may not be in a form suitable for testing.

Defining a research problem properly and clearly is a crucial part of a research study and must in no case be accomplished hurriedly. However, in practice this is frequently overlooked which causes a lot of problems later on. Hence, the research problem should be defined in a systematic manner, giving due weightage to all relating points. The technique for the purpose involves the undertaking of the following steps generally one after the other: (i) statement of the problem in a general way; (ii) understanding the nature of the problem; (iii) surveying the available literature; (iv) developing the ideas through discussions; and (v) rephrasing the research problem into a working proposition.

A brief description of all these points will be helpful.

(i) **Statement of the problem in a general way:** First of all the problem should be stated in a broad general way, keeping in view either some practical concern or some scientific or intellectual interest. For this purpose, the researcher must immerse himself thoroughly in the subject matter concerning which he wishes to pose a problem. In case of social research, it is considered advisable to do some field observation and as such the researcher may undertake some sort of preliminary survey or what is often called *pilot survey*. Then the researcher can himself state the problem or he can seek the guidance of the guide or the subject expert in accomplishing this task. Often, the guide puts forth the problem in general terms, and it is then up to the researcher to narrow it down and phrase the problem in operational terms. In case there is some directive from an organisational authority, the problem then can be stated accordingly. The problem stated in a broad general way may contain various ambiguities which must be resolved by cool thinking and rethinking over the problem. At the same time the feasibility of a particular solution has to be considered and the same should be kept in view while stating the problem.

(ii) **Understanding the nature of the problem:** The next step in defining the problem is to understand its origin and nature clearly. The best way of understanding the problem is to discuss it with those who first raised it in order to find out how the problem originally came about and with what objectives in view. If the researcher has stated the problem himself, he should consider once again all those points that induced him to make a general statement concerning the problem. For a better understanding of the nature of the problem involved, he can enter into discussion with those who have a good knowledge of the problem concerne or similar other problems. The researcher should also keep in view the environment with which the problem is to be studied and understood.

(iii) **Surveying the available literature:** All available literature concerning the problem hand must necessarily be surveyed and examined before a definition of the research proble is given. This means that the researcher must be well-conversant with relevant theories i the field, reports and records as also all other relevant literature. He must devote sufficie time in reviewing of research already undertaken on related problems. This is done to fin out what data and other materials, if any, is available for operational purposes. This woul also help a researcher to know if there are certain gaps in the theories, or whether th existing theories applicable to the problem under study are inconsistent with each other, whether the findings of the different studies do not follow a pattern consistent with th theoretical expectations and so on. All this will enable a researcher to take new strides i the field for furtherance of knowledge i.e., he can move up starting from the existin premise. Studies on related problems are useful for indicating the type of difficulties tha may be encountered in the present study as also the possible analytical shortcomings. A times such studies may also suggest useful and even new lines of approach to the prese problem.

(iv) **Developing the ideas through discussions:** Discussion concerning a problem ofte produces useful information. Various new ideas can be developed through such an exercise Hence, a researcher must discuss his problem with his colleagues and others who hav enough experience in the same area or in working on similar problems. This is quite ofte known as an *experience survey*. People with rich experience are in a position to enlighte the researcher on different aspects of his proposed study and their advice and comment are usually invaluable to the researcher. They help him sharpen his focus of attention o specific aspects within the field. Discussions with such persons should not only be confine to the formulation of the specific problem at hand, but should also be concerned with th general approach to the given problem, techniques that might be used, possible solutions etc.

(v) **Rephrasing the research problem:** Finally, the researcher must sit to rephrase th research problem into a working proposition. Once the nature of the problem has bee clearly understood, the environment (within which the problem has got to be studied) ha been defined, discussions over the problem have taken place and the available literatur has been surveyed and examined, rephrasing the problem into analytical or operationa terms is not a difficult task. Through rephrasing, the researcher puts the research proble in as specific terms as possible so that it may become operationally viable and may help i the development of working hypothesis.*

---

* Working hypotheses are a set of suggested tentative solutions of explanations of a research problem which may or may not be the real solutions. The task of research is to test and establish such hypothesis. Hypothesis should be clearly an precisely stated in simple terms, they should be testable, limited in scope and should state relationship between variables They should be amenable to testing within a reasonable time and should be consistent with most of the known fact (Testing of hypothesis has been dealt later in the book).

In addition to what has been stated above, the following points must also be observed while defining a research problem:

(a) Technical terms and words or phrases, with special meanings used in the statement of the problem, should be clearly defined.

(b) Basic assumptions or postulates (if any) relating to the research problem should be clearly stated.

(c) A straight forward statement of the value of the investigation (i.e., the criteria for the selection of the problem) should be provided.

(d) The suitability of the time-period and the sources of data available must also be considered by the researcher in defining the problem.

(e) The scope of the investigation or the limits within which the problem is to be studied must be mentioned explicitly in defining a research problem.

2.5 AN ILLUSTRATION

The technique of defining a problem outlined above can be illustrated for better understanding by taking an example as under:

Let us suppose that a research problem in a broad general way is as follows:

> "Why is productivity in Japan so much higher than in India"?
>
> In this form the question has a number of ambiguities such as: What sort of productivity is being referred to? With what industries the same is related? With what period of time the productivity is being talked about? In view of all such ambiguities the given statement or the question is much too general to be amenable to analysis. Rethinking and discussions about the problem may result in narrowing down the question to:
>
> "What factors were responsible for the higher labour productivity of Japan's manufacturing industries during the decade 1971 to 1980 relative to India's manufacturing industries?"
>
> This latter version of the problem is definitely an improvement over its earlier version for the various ambiguities have been removed to the extent possible. Further rethinking and rephrasing might place the problem on a still better operational basis as shown below:
>
> "To what extent did labour productivity in 1971 to 1980 in Japan exceed that of India in respect of 15 selected manufacturing industries? What factors were responsible for the productivity differentials between the two countries by industries?"

With this sort of formulation, the various terms involved such as 'labour productivity', 'productivity differentials', etc. must be explained clearly. The researcher must also see that the necessary data is available. In case the data for one or more industries selected is not available for the concerning time-period, then the said industry or industries will have to be substituted by other industry or industries. The suitability of the time-period must also be examined. Thus, all relevant factors must be considered by a researcher before finally defining a research problem.

2.6 CONCLUSION

We may conclude by saying that the task of defining a research problem, very often, follows sequential pattern—the problem is stated in a general way, the ambiguities are resolved, thinking an rethinking process results in a more specific formulation of the problem so that it may be a realist one in terms of the available data and resources and is also analytically meaningful. All this results a well defined research problem that is not only meaningful from an operational point of view, but equally capable of paving the way for the development of working hypotheses and for means o solving the problem itself.

PROBLEMS

1. Describe fully the techniques of defining a research problem.
2. What is research problem? Define the main issues which should receive the attention of the researcher in formulating the research problem. Give suitable examples to elucidate your points.
3. How do you define a research problem? Give three examples to illustrate your answer.
4. What is the necessity of defining a research problem? Explain.
5. Write short notes on:
    (a) Experience survey;
    (b) Pilot survey;
    (c) Components of a research problem;
    (d) Rephrasing the research problem.
6. "The task of defining the research problem often follows a sequential pattern". Explain.
7. "Knowing what data are available often serves to narrow down the problem itself as well as the technique that might be used." Explain the underlying idea in this statement in the context of defining a research problem.
8. Write a comprehensive note on the "Task of defining a research problem".

3

Chapter

Research Design

3.1 MEANING OF RESEARCH DESIGN

The formidable problem that follows the task of defining the research problem is the preparation of a design of the research project, popularly known as "research design". A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure. In fact, the research design is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. As such the design includes an outline of what the researcher will do from writing the hypothesis and its operational implications to the final analysis of data. More explicitly, the design decisions happen to be in respect of:

(i) What is the study about?
(ii) Why is the study being made?
(iii) Where will the study be carried out?
(iv) What type of data is required?
(v) Where can the required data be found?
(vi) What periods of time will the study include?
(vii) What will be the sample design?
(viii) What techniques of data collection will be used?
(ix) How will the data be analysed?
(x) In what style will the report be prepared?

Keeping in view the above stated design decisions, one may split the overall research design into the following parts:

(a) *The sampling design* which deals with the method of selecting items to be observed for the given study.

(b) *The observational design* which relates to the conditions under which the observations are to be made.

(c) *The statistical design* which concerns the question of how many items are to be observed and how the information and data gathered is to be analysed.

(d) *The operational design* which deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

From what has been stated above, we can state the important features of a research design as under:

(i) It is a plan that specifies the sources and types of information relevant to the research problem.

(ii) It is a strategy specifying which approach will be used for gathering and analysing the data.

(iii) It also includes the time and cost budgets since most studies are done under these two constraints.

In brief, research design must, at least, contain (a) a clear statement of the research problem; (b) procedures and techniques to be used for gathering information; (c) the population to be studied; and (d) methods to be used in processing and analysing data.

3.2 NEED FOR RESEARCH DESIGN

Research design is needed because it facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money. Just as for better, economical and attractive construction of a house, we need a blueprint (or what is commonly called the map of the house) well thought out and prepared by an expert architect, similarly, we need a research design or a plan in advance of data collection and analysis for our research project. Research design stands for advance planning of the methods to be adopted for collecting the relevant data and the techniques to be used in their analysis, keeping in view the objective of the research and the availability of staff, time and money. Preparation of the research design should be done with great care as any error in it may upset the entire project. Research design, in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work.

Even then the need for a well thought out research design is at times not realised by many. The importance which this problem deserves is not given to it. As a result many researches do not serve the purpose for which they are undertaken. In fact, they may even give misleading conclusions. Thoughtlessness in designing the research project may result in rendering the research exercise futile. It is, therefore, imperative that an efficient and appropriate design must be prepared before starting research operations. The design helps the researcher to organize his ideas in a form whereby it will be possible for him to look for flaws and inadequacies. Such a design can even be given to others for their comments and critical evaluation. In the absence of such a course of action, it will be difficult for the critic to provide a comprehensive review of the proposed study.

3.3 FEATURES OF A GOOD DESIGN

good design is often characterised by adjectives like flexible, appropriate, efficient, economical and on. Generally, the design which minimises bias and maximises the reliability of the data collected nd analysed is considered a good design. The design which gives the smallest experimental error is upposed to be the best design in many investigations. Similarly, a design which yields maximal nformation and provides an opportunity for considering many different aspects of a problem is onsidered most appropriate and efficient design in respect of many research problems. Thus, the uestion of good design is related to the purpose or objective of the research problem and also with e nature of the problem to be studied. A design may be quite suitable in one case, but may be found anting in one respect or the other in the context of some other research problem. One single design annot serve the purpose of all types of research problems.

A research design appropriate for a particular research problem, usually involves the consideration f the following factors:

(i) the means of obtaining information;
(ii) the availability and skills of the researcher and his staff, if any;
(iii) the objective of the problem to be studied;
(iv) the nature of the problem to be studied; and
(v) the availability of time and money for the research work.

If the research study happens to be an exploratory or a formulative one, wherein the major mphasis is on discovery of ideas and insights, the research design most appropriate must be flexible nough to permit the consideration of many different aspects of a phenomenon. But when the purpose f a study is accurate description of a situation or of an association between variables (or in what are alled the descriptive studies), accuracy becomes a major consideration and a research design which inimises bias and maximises the reliability of the evidence collected is considered a good design. tudies involving the testing of a hypothesis of a causal relationship between variables require a esign which will permit inferences about casuality in addition to the minimisation of bias and aximisation of reliability. But in practice it is the most difficult task to put a particular study in a articular group, for a given research may have in it elements of two or more of the functions of ifferent studies. It is only on the basis of its primary function that a study can be categorised either s an exploratory or descriptive or hypothesis-testing study and accordingly the choice of a research esign may be made in case of a particular study. Besides, the availability of time, money, skills of the esearch staff and the means of obtaining the information must be given due weightage while working ut the relevant details of the research design such as experimental design, survey design, sample esign and the like.

3.4 IMPORTANT CONCEPTS RELATING TO RESEARCH DESIGN

Before describing the different research designs, it will be appropriate to explain the various concepts elating to designs so that these may be better and easily understood.

3.4.1 Dependent and Independent Variables

A concept which can take on different quantitative values is called a *variable*. As such the concep like weight, height, income are all examples of variables. Qualitative phenomena (or the attribute are also quantified on the basis of the presence or absence of the concerning attribute(s). Phenomen which can take on quantitatively different values even in decimal points are called 'continuou variables'.* But all variables are not continuous. If they can only be expressed in integral values, the are non-continuous variables or in statistical language 'discrete variables'.** Age is an example c continuous variable, but the number of children is an example of non-continuous variable. If on variable depends upon or is a consequence of the other variable, it is termed as a dependent variabl and the variable that is antecedent to the dependent variable it is termed as an independent variabl For instance, if we say that height depends upon age, then height is a dependent variable and age an independent variable. Further, if in addition to being dependent upon age, height also depends upo the individual's sex, then height is a dependent variable and age and sex are independent variable Similarly, readymade films and lectures are examples of independent variables, whereas behavioura changes, occurring as a result of the environmental manipulations, are examples of dependent variable

3.4.2 Extraneous Variable

Independent variables that are not related to the purpose of the study, but may affect the dependen variable are termed as extraneous variables. Suppose the researcher wants to test the hypothesi that there is a relationship between children's gains in social studies achievement and their sel concepts. In this case self-concept is an independent variable and social studies achievement is dependent variable. Intelligence may as well affect the social studies achievement, but since it is nc related to the purpose of the study undertaken by the researcher, it will be termed as an extraneou variable. Whatever effect is noticed on dependent variable as a result of extraneous variable(s) i technically described as an 'experimental error'. A study must always be so designed that *the effec upon the dependent variable is attributed entirely to the independent variable(s), and not t some extraneous variable or variables.*

3.4.3 Control

One important characteristic of a good research design is to minimise the influence or effect c extraneous variable(s). The technical term 'control' is used when we design the study minimising th effects of extraneous independent variables. In experimental researches, the term 'control' is use to refer to restrain experimental conditions.

3.4.4 Confounded Relationship

When the dependent variable is not free from the influence of extraneous variable(s), the relationshi between the dependent and independent variables is said to be confounded by an extraneou variable(s).

---

* A continuous variable is that which can assume any numerical value within a specific range.

** A variable for which the individual values fall on the scale only with distinct gaps is called a discrete variable.

.4.5 Research Hypothesis

hen a prediction or a hypothesised relationship is to be tested by scientific methods, it is termed as esearch hypothesis. The research hypothesis is a predictive statement that relates an independent ariable to a dependent variable. Usually, a research hypothesis must contain, at least, one independent nd one dependent variable. Predictive statements which are not to be objectively verified or the elationships that are assumed but not to be tested, are not termed research hypothesis.

3.4.6 Experimental and Non-Experimental Hypothesis-Testing Research

When the purpose of research is to test a research hypothesis, it is termed as hypothesis-testing research. It can be of the experimental design or of the non-experimental design. Research in which the independent variable is manipulated is termed 'experimental hypothesis-testing research' and a research in which an independent variable is not manipulated is called 'non-experimental hypothesis-testing research'. For instance, suppose a researcher wants to study whether intelligence affects reading ability for a group of students and for this purpose he randomly selects 50 students and tests their intelligence and reading ability by calculating the coefficient of correlation between the two sets of scores. This is an example of non-experimental hypothesis-testing research because herein the independent variable, intelligence, is not manipulated. But now suppose that our researcher randomly selects 50 students from a group of students who are to take a course in statistics and then divides them into two groups by randomly assigning 25 to Group A, the usual studies programme, and 25 to Group B, the special studies programme. At the end of the course, he administers a test to each group in order to judge the effectiveness of the training programme on the student's performance-level. This is an example of experimental hypothesis-testing research because in this case the independent variable, viz., the type of training programme, is manipulated.

3.4.7 Experimental and Control Groups

In an experimental hypothesis-testing research when a group is exposed to usual conditions, it is termed a 'control group', but when the group is exposed to some novel or special condition, it is termed as 'experimental group'. In the above illustration, the Group A can be called a control group and the Group B an experimental group. If both groups A and B are exposed to special studies programme, then both groups would be termed 'experimental groups.' It is possible to design studies which include only experimental groups or studies.

3.4.8 Treatments

The different conditions under which experimental and control groups are put are usually referred to as 'treatments'. In the illustration taken above, the two treatments are the usual studies programme and the special studies programme. Similarly, if we want to determine through an experiment the comparative impact of three varieties of fertilizers on the yield of wheat, in that case the three varieties of fertilizers will be treated as three treatments.

3.4.9 Experiment

The process of examining the truth of a statistical hypothesis, relating to some research problem, known as an experiment. For example, we can conduct an experiment to examine the usefulness a certain newly developed drug. Experiments can be of two types viz., absolute experiment a comparative experiment. If we want to determine the impact of a fertilizer on the yield of a crop, is a case of absolute experiment; but if we want to determine the impact of one fertilizer as compar to the impact of some other fertilizer, our experiment then will be termed as a comparative experime Often, we undertake comparative experiments when we talk about designs of experiments.

3.4.10 Experimental Unit(s)

The pre-determined plots or the blocks, where different treatments are used, are known experimental units. Such experimental units must be selected (defined) very carefully.

3.5 DIFFERENT RESEARCH DESIGNS

Different research designs can be categorized as below:

We take up each category separately.

3.5.1 Research Design in Case of Exploratory Research Studies

Exploratory research studies are also termed as formulative research studies. The main purpose such studies is that of formulating a problem for more precise investigation or of developing t working hypothesis from an operational point of view. The major emphasis in such studies is on t discovery of ideas and insights. As such the research design appropriate for such studies must flexible enough to provide opportunity for considering different aspects of a problem under stu Inbuilt flexibility in research design is needed because the research problem, broadly defined initial is transformed into one with more precise meaning in exploratory studies, which infact may necessit changes in the research procedure for gathering relevant data. Generally, the following three metho in the context of research design for such studies are talked about: (a) the survey of concerni literature; (b) the experience survey and (c) the analysis of 'insight-stimulating' examples.

The survey of concerning literature happens to be the most simple and fruitful method formulating precisely the research problem or developing hypothesis. Hypothesis stated by earli workers may be reviewed and their usefulness be evaluated as a basis for further research. It m also be considered whether the already stated hypothesis suggest new hypothesis. In this way t researcher should review and build upon the work already done by others, but in cases whe hypothesis have not yet been formulated, his task is to review the available material for deriving t relevant hypothesis from it.

Besides, the bibliographical survey of studies, already made in one's area of interest may as w be made by the researcher for precisely formulating the problem. He should also make an attempt apply concepts and develop theories in different research contexts to the area in which he is hims working. Sometimes the works of creative writers also provide a fertile ground for hypothes formulation and as such may be looked into by the researcher.

Experience survey means the survey of people who have had practical experience with the roblem to be studied. The object of such a survey is to obtain insight into the relationships between ariables and new ideas relating to the research problem. For such a survey people who are competent nd can contribute new ideas may be carefully selected as respondents to ensure a representation of ifferent types of experience. The respondents so selected may then be interviewed by the investigator. he researcher must prepare an interview schedule for the systematic questioning of informants. ut the interview must ensure flexibility in the sense that the respondents should be allowed to raise ssues and questions which the investigator has not previously considered. Generally, the experience-ollecting interview is likely to be long and may last for few hours. Hence, it is often considered esirable to send a copy of the questions to be discussed to the respondents well in advance. This will lso give an opportunity to the respondents for doing some advance thinking over the various issues nvolved so that, at the time of interview, they may be able to contribute effectively. Thus, an experience urvey may enable the researcher to define the problem more concisely and help in the formulation f the research hypothesis. This survey may as well provide information about the practical possibilities or doing different types of research.

Analysis of 'insight-stimulating' examples is also a fruitful method for suggesting hypothesis or research. It is particularly suitable in areas where there is little experience to serve as a guide. This method consists of the intensive study of selected instances of the phenomenon in which one is nterested. For this purpose the existing records, if any, may be examined, the unstructured interviewing nay take place, or some other approach may be adopted. Attitude of the investigator, the intensity of he study and the ability of the researcher to draw together diverse information into a unified nterpretation are the main features which make this method an appropriate procedure for evoking nsights.

Now, what sort of examples are to be selected and studied? There is no clear cut answer to it. Experience indicates that for particular problems certain types of instances are more appropriate han others. One can mention few examples of 'insight-stimulating' cases such as the reactions of strangers, the reactions of marginal individuals, the study of individuals who are in transition from one stage to another, the reactions of individuals from different social strata and the like. In general, cases that provide sharp contrasts or have striking features are considered relatively more useful while adopting this method of hypothesis formulation.

Thus, in an exploratory of formulative research study which merely leads to insights or hypothesis, whatever method or research design outlined above is adopted, the only thing essential is that it must continue to remain flexible so that many different facets of a problem may be considered as and when they arise and come to the notice of the researcher.

3.5.2 Research Design in Case of Descriptive and Diagnostic Research Studies

Descriptive research studies are those studies which are concerned with describing the characteristics of a particular individual, or of a group, whereas diagnostic research studies determine the frequency with which something occurs or its association with something else. The studies concerning whether certain variables are associated are examples of diagnostic research studies. As against this, studies concerned with specific predictions, with narration of facts and characteristics concerning individual, group or situation are all examples of descriptive research studies. Most of the social research comes under this category. From the point of view of the research design, the descriptive as well a diagnostic studies share common requirements and as such we may group together these two typ of research studies. In descriptive as well as in diagnostic studies, the researcher must be able t define clearly, what he wants to measure and must find adequate methods for measuring it alon with a clear cut definition of 'population' he wants to study. Since the aim is to obtain complete an accurate information in the said studies, the procedure to be used must be carefully planned. Th research design must make enough provision for protection against bias and must maximise reliabilit with due concern for the economical completion of the research study. The design in such studi must be rigid and not flexible and must focus attention on the following:

(a) Formulating the objective of the study (what the study is about and why is it being made
(b) Designing the methods of data collection (what techniques of gathering data will be adopted
(c) Selecting the sample (how much material will be needed?)
(d) Collecting the data (where can the required data be found and with what time period shoul the data be related?)
(e) Processing and analysing the data.
(f) Reporting the findings.

In a descriptive/diagnostic study the first step is to specify the objectives with sufficient precisio to ensure that the data collected is relevant. If this is not done carefully, the study may not provide th desired information.

Then comes the question of selecting the methods by which the data is to be obtained. In othe words, techniques for collecting the information must be devised. Several methods (viz., observatio questionnaires, interviewing, examination of records, etc.), with their merits and limitations, are availabl for the purpose and the researcher may use one or more of these methods which have been discusse in detail in the later chapters. While designing data-collection procedure, adequate safeguards agains bias and unreliability must be ensured. Whichever method is selected, questions must be well examine and be made unambiguous; interviewers must be instructed not to express their own opinion; observer must be trained so that they uniformly record a given item of behaviour. It is always desirable to pre test the data collection instruments before they are finally used for the study purposes. In othe words, we can say that "*structured instruments*" are used in such studies.

In most of the descriptive/diagnostic studies the researcher takes out sample(s) and then wishe to make statements about the population on the basis of the sample analysis or analyses. More ofte than not, sample has to be designed. Different sample designs have been discussed in detail in separate chapter in this book. Here, we may only mention that the problem of designing sample should be tackled in such a fashion that the samples may yield accurate information with a minimu amount of research effort. Usually one or more forms of probability sampling, or what is ofte described as random sampling, are used.

To obtain data free from errors introduced by those responsible for collecting them, it is necessar to supervise closely the staff of field workers as they collect and record information. Checks may b set up to ensure that the data collecting staff perform their duty honestly and without prejudice Collected data should be examined for its completeness, comprehensibility, consistency and reliabilit

The data collected must be processed and analysed. This includes steps like coding the interview eplies, observations, etc.; tabulating the data; and performing several statistical computations. To ne extent possible, the processing and analysing procedure should be planned in detail before actual vork is started. This will prove economical in the sense that the researcher may avoid unnecessary abour such as preparing tables for which he later finds he has no use or on the other hand, re-doing ome tables because he failed to include relevant data. Coding should be done carefully to avoid rror in coding and for this purpose the reliability of coders needs to be checked. Similarly, the ccuracy of tabulation may be checked by having a sample of the tables re-done. In case of mechanical abulation the material (i.e., the collected data or information) must be entered on appropriate cards vhich is usually done by punching holes corresponding to a given code. The accuracy of punching is o be checked and ensured. Finally, statistical computations are needed and as such averages, ercentages and various coefficients must be worked out. Probability and sampling analysis may as vell be used. The appropriate statistical operations, along with the use of appropriate tests of ignificance should be carried out to safeguard the drawing of conclusions concerning the study.

Last of all comes the question of reporting the findings. This is the task of communicating the indings to others and the researcher must do it in an efficient manner. The layout of the report needs o be well planned so that all things relating to the research study may be well presented in simple and ffective style.

Thus, the research design in case of descriptive/diagnostic studies is a comparative design throwing ight on all points narrated above and must be prepared keeping in view the objective(s) of the study nd the resources available. However, it must ensure the minimisation of bias and maximisation of eliability of the evidence collected. The said design can be appropriately referred to as a *survey lesign* since it takes into account all the steps involved in a survey concerning a phenomenon to be tudied.

The difference between research designs in respect of the above two types of research studies an be conveniently summarised in tabular form as under:

Table 3.1

Research Design	*Type of Study*	
	Exploratory of Formulative	*Descriptive/Diagnostic*
Overall design	Flexible design (design must provide opportunity for considering different aspects of the problem)	Rigid design (design must make enough provision for protection against bias and must maximise reliability)
(i) Sampling design	Non-probability sampling design (purposive or judgement sampling)	Probability sampling design (random sampling)
(ii) Statistical design	No pre-planned design for analysis	Pre-planned design for analysis
(iii) Observational design	Unstructured instruments for collection of data	Structured or well thought out instruments for collection of data
(iv) Operational design	No fixed decisions about the operational procedures	Advanced decisions about operational procedures.

3.5.3 Research Design in Case of Hypothesis-Testing Research Studies

Hypothesis-testing research studies (generally known as experimental studies) are those where t researcher tests the hypothesis of casual relationships between variables. Such studies requ procedures that will not only reduce bias and increase reliability, but will permit drawing inferenc about casuality. Usually experiments meet this requirement. Hence, when we talk of research desi in such studies, we often mean the design of experiments.

Professor R.A. Fisher's name is associated with experimental designs. Beginning of such desig was made by him when he was working at Rothamsted Experimental Station (Centre for Agricultu Research in England). As such the study of experimental designs has its origin in agricultural researc Professor Fisher found that by dividing agricultural fields or plots into different blocks and then conducting experiments in each of these blocks, whatever information is collected and inferenc drawn from them, happens to be more reliable. This fact inspired him to develop certain experimen designs for testing hypothesis concerning scientific investigations. Today, the experimental desig are being used in researches relating to phenomena of several disciplines. Since experimental desig originated in the context of agricultural operations, we still use, though in a technical sense, seve terms of agriculture (such as treatment, yield, plot, block etc.) in experimental designs.

3.6 BASIC PRINCIPLES OF EXPERIMENTAL DESIGNS

Professor Fisher has enumerated three principles of experimental designs: The Principle of Replicatic The Principle of Randomization; and The Principle of Local Control.

According to the *Principle of Replication*, the experiment should be repeated more than onc Thus, each treatment is applied in many experimental units instead of one. By doing so the statistic accuracy of the experiments is increased. For example, suppose we are to examine the effect of tv varieties of rice. For this purpose we may divide the field into two parts and grow one variety in o part and the other variety in the other part. We can then compare the yield of the two parts and dra conclusion on that basis. But if we are to apply the principle of replication to this experiment, then v first divide the field into several parts, grow one variety in half of these parts and the other variety the remaining parts. We can then collect the data of yield of the two varieties and draw conclusion l comparing the same. The result so obtained will be more reliable in comparison to the conclusion v draw without applying the principle of replication. The entire experiment can even be repeat several times for better results. Conceptually replication does not present any difficulty, b computationally it does. For example, if an experiment requiring a two-way analysis of variance replicated, it will then require a three-way analysis of variance since replication itself may be source of variation in the data. However, it should be remembered that replication is introduced order to increase the precision of a study; that is to say, to increase the accuracy with which the ma effects and interactions can be estimated.

The *Principle of Randomization* provides protection when we conduct an experiment again the effect of extraneous factors by randomization. In other words, this principle indicates that v should design or plan the experiment in such a way that the variations caused by extraneous facto can all be combined under the general heading of "chance." For instance, if we grow one variety of rice, say, in the first half of the parts of a field and the other variety is grown in the other half, then it is just possible that the soil fertility may be different in the first half in comparison to the other half. If this is so, our results would not be realistic. In such a situation, we may assign the variety of rice to be grown in different parts of the field on the basis of some random sampling technique i.e., we may apply randomization principle and protect ourselves against the effects of the extraneous factors (soil fertility differences in the given case). As such, through the application of the principle of randomization, we can have a better estimate of the experimental error.

The *Principle of Local Control* is another important principle of experimental designs. Under it the extraneous factor, the known source of variability, is made to vary deliberately over as wide range as necessary and this needs to be done in such a way that the variability it causes can be measured and hence eliminated from the experimental error. This means that we should plan the experiment in a manner that we can perform a two-way analysis of variance, in which the total variability of the data is divided into three components attributed to treatments (varieties of rice in our case), the extraneous factor (soil fertility in our case) and experimental error.* In other words, according to the principle of local control, we first divide the field into several homogeneous parts, known as blocks, and then each such block is divided into parts equal to the number of treatments. Then the treatments are randomly assigned to these parts of a block. Dividing the field into several homogeneous parts is known as 'blocking'. In general, blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of a two-way analysis of variance. In brief, through the principle of local control we can eliminate the variability due to extraneous factor(s) from the experimental error.

3.7 IMPORTANT EXPERIMENTAL DESIGNS

Experimental design refers to the framework or structure of an experiment and as such there are several experimental designs. We can classify experimental designs into two broad categories, viz., informal experimental designs and formal experimental designs. Informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental designs offer relatively more control and use precise statistical procedures for analysis. Important experimental designs are as follows:

(a) Informal experimental designs:
  (i) Before-and-after without control design.
  (ii) After-only with control design.
  (iii) Before-and-after with control design.

(b) Formal experimental designs:
  (i) Completely randomized design (C.R. Design).
  (ii) Randomized block design (R.B. Design).
  (iii) Latin square design (L.S. Design).
  (iv) Factorial designs.

---

*See Chapter Analysis of Variance for details.

We may briefly deal with each of the above stated informal as well as formal experimenta designs.

3.7.1 Before-and-After Without Control Design

In such a design a single test group or area is selected and the dependent variable is measured befor the introduction of the treatment. The treatment is then introduced and the dependent variable i measured again after the treatment has been introduced. The effect of the treatment would be equa to the level of the phenomenon after the treatment minùs the level of the phenomenon before th treatment. The design can be represented thus:

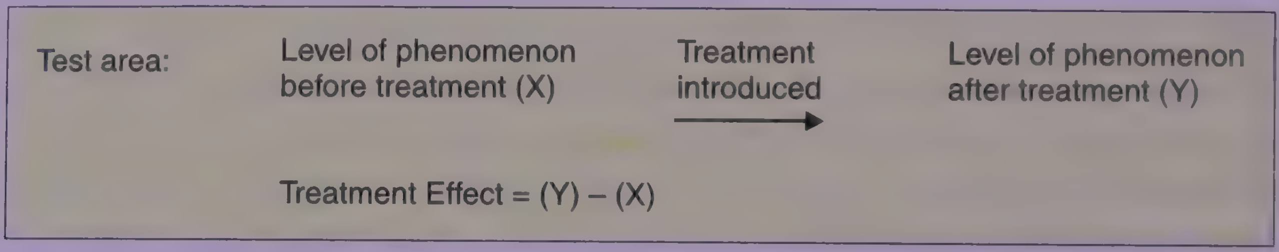

Test area:
Level of phenomenon before treatment (X)
Treatment introduced
Level of phenomenon after treatment (Y)
Treatment Effect = (Y) – (X)

Fig. 3.1

The main difficulty of such a design is that with the passage of time considerable extraneou variations may be there in its treatment effect.

3.7.2 After-Only with Control Design

In this design two groups or areas (test area and control area) are selected and the treatment i introduced into the test area only. The dependent variable is then measured in both the areas at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control area from its value in the test area. This can be exhibited in the following form:

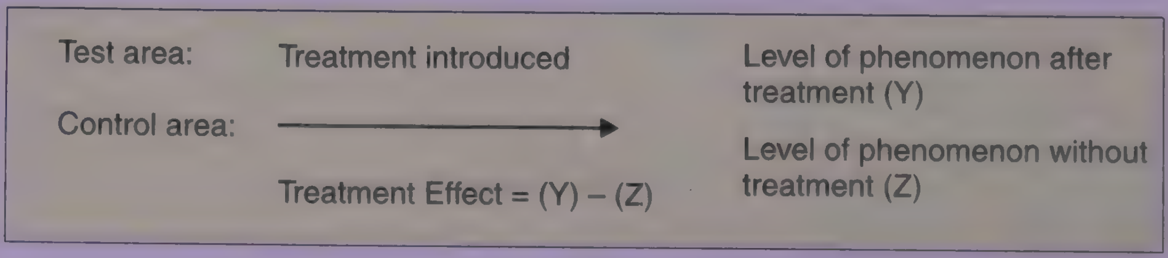

Test area:
Control area:
Treatment introduced
Level of phenomenon after treatment (Y)
Level of phenomenon without treatment (Z)
Treatment Effect = (Y) – (Z)

Fig. 3.2

The basic assumption in such a design is that the two areas are identical with respect to thei behaviour towards the phenomenon considered. If this assumption is not true, there is the possibilit of extraneous variation entering into the treatment effect. However, data can be collected in such design without the introduction of problems with the passage of time. In this respect the design i superior to before-and-after without control design.

.7.3 Before-and-After with Control Design

n this design two areas are selected and the dependent variable is measured in both the areas for an lentical time-period before the treatment. The treatment is then introduced into the test area only, nd the dependent variable is measured in both for an identical time-period after the introduction of ne treatment. The treatment effect is determined by subtracting the change in the dependent variable n the control area from the change in the dependent variable in test area. This design can be shown n this way:

	Time Period I		Time Period II
Test area:	Level of phenomenon before treatment (X)	Treatment introduced →	Level of phenomenon after treatment (Y)
Control area:	Level of phenomenon without treatment (A)		Level of phenomenon without treatment (Z)

Treatment Effect = (Y – X) – (Z – A)

Fig. 3.3

This design is superior to the above two designs for the simple reason that it avoids extraneous variation resulting both from the passage of time and from non-comparability of the test and control areas. But at times, due to lack of historical data, time or a comparable control area, we should prefer to select one of the first two informal designs stated above.

3.7.4 Completely Randomized Design (C.R. Design)

Involves only two principles viz., the principle of replication and the principle of randomization of experimental designs. It is the simplest possible design and its procedure of analysis is also easier. The essential characteristic of the design is that subjects are randomly assigned to experimental treatments (or vice-versa). For instance, if we have 10 subjects and if we wish to test 5 under treatment A and 5 under treatment B, the randomization process gives every possible group of 5 subjects selected from a set of 10 an equal opportunity of being assigned to treatment A and treatment B. One-way analysis of variance (or one-way ANOVA)* is used to analyse such a design. Even unequal replications can also work in this design. It provides maximum number of degrees of freedom to the error. Such a design is generally used when experimental areas happen to be homogeneous. Technically, when all the variations due to uncontrolled extraneous factors are included under the heading of chance variation, we refer to the design of experiment as C.R. design.

We can present a brief description of the two forms of such a design as given in Fig. 3.4.

(i) **Two-group simple randomized experimental design:** In a two-group simple randomized design, first of all the population is defined and then from the population a sample is selected randomly. Further, requirement of this design is that items, after being selected randomly from the population, be randomly assigned to the experimental and control groups (suc random assignment of items to two groups is technically described as principle randomization). Thus, this design yields two groups as representatives of the population. I a diagram form this design can be shown in this way:

---

* See Chapter 12 for one-way ANOVA technique.

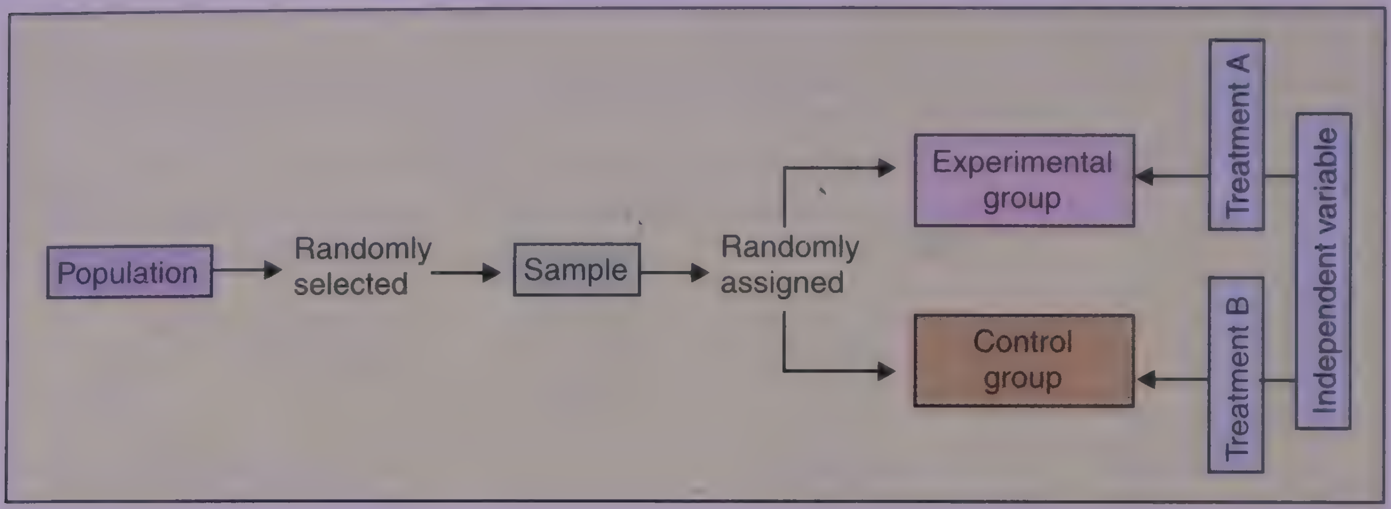


Fig. 3.4: Two-group simple randomized experimental design (in diagram form)

Since in the sample randomized design the elements constituting the sample are randoml drawn from the same population and randomly assigned to the experimental and contro groups, it becomes possible to draw conclusions on the basis of samples applicable for th population. The two groups (experimental and control groups) of such a design are give different treatments of the independent variable. This design of experiment is quite commo in research studies concerning behavioural sciences. The merit of such a design is that it i simple and randomizes the differences among the sample items. But the limitation of it i that the individual differences among those conducting the treatments are not eliminate i.e., it does not control the extraneous variable and as such the result of the experiment ma not depict a correct picture. This can be illustrated by taking an example. Suppose th researcher wants to compare two groups of students who have been randomly selecte and randomly assigned. Two different treatments viz., the usual training and the specialise training are being given to the two groups. The researcher hypothesis greater gains for th group receiving specialised training. To determine this, he tests each group before and afte the training, and then compares the amount of gain for the two groups to accept or rejec his hypothesis. This is an illustration of the two-groups randomized design, wherein individua differences among students are being randomized. But this does not control the differentia effects of the extraneous independent variables (in this case, the individual difference among those conducting the training programme).

(ii) Random replications design: The limitation of the two-group randomized design is usuall eliminated within the random replications design. In the illustration just cited above, th *teacher differences* on the dependent variable were ignored, i.e., the extraneous variabl was not controlled. But in a random replications design, the effect of such differences ar minimised (or reduced) by providing a number of repetitions for each treatment. Eac repetition is technically called a 'replication'. Random replication design serves two purpose viz., it provides controls for the differential effects of the extraneous independent variables and secondly, it randomizes any individual differences among those conducting the treatments. Diagrammatically we can illustrate the random replications design thus: (Fig. 3.5).

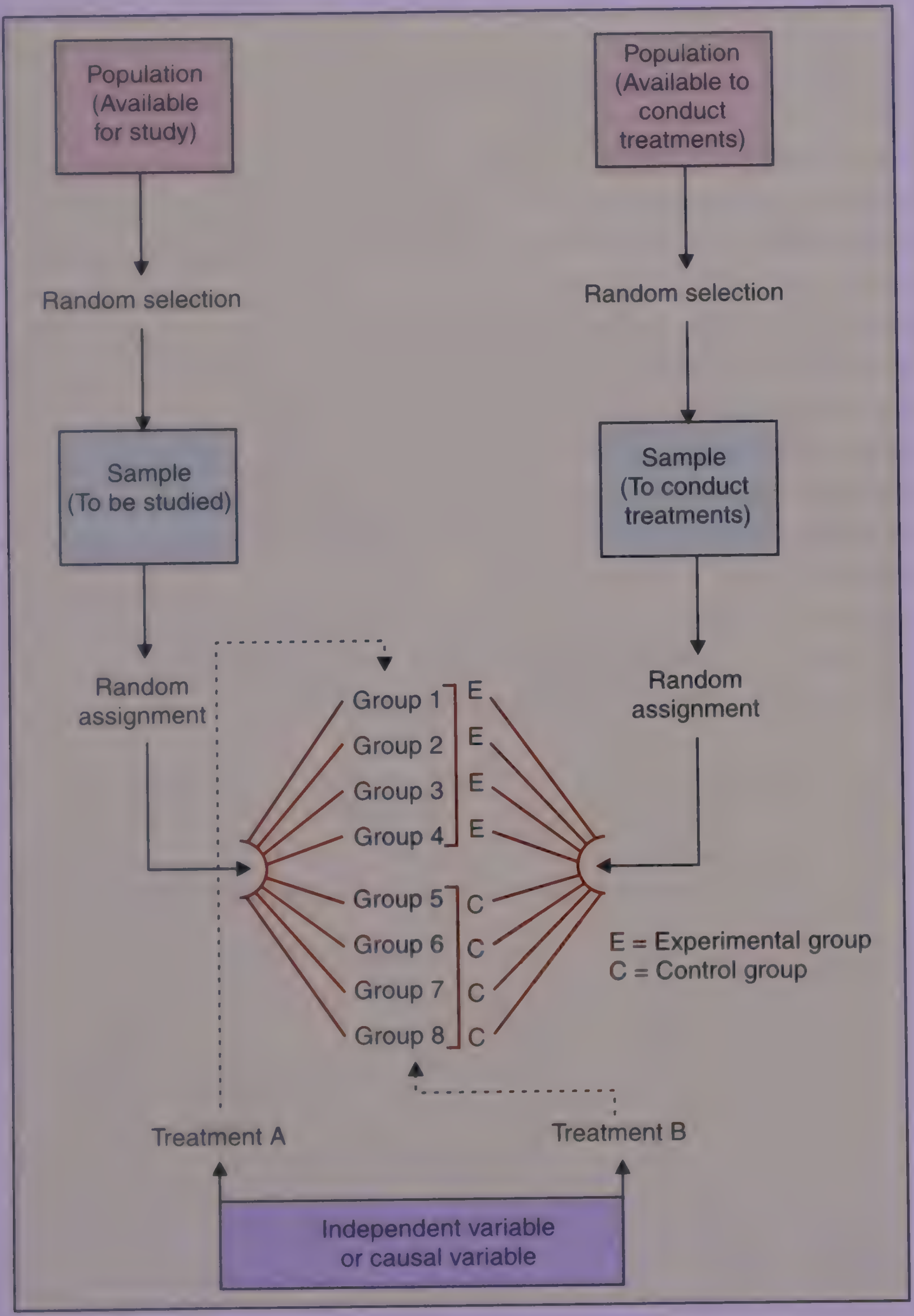

Population (Available for study)
Random selection
Sample (To be studied)
Random assignment
Population (Available to conduct treatments)
Random selection
Sample (To conduct treatments)
Random assignment
Group 1 E
Group 2 E
Group 3 E
Group 4 E
Group 5 C
Group 6 C
Group 7 C
Group 8 C
E = Experimental group
C = Control group
Treatment A
Treatment B
Independent variable or causal variable

Fig. 3.5: Random replication design (in diagram form)

From the diagram it is clear that there are two populations in the replication design. The sample is taken randomly from the population available for study and is randomly assigned to, say, four experimental and four control groups. Similarly, sample is taken randomly from the population available to conduct experiments (because of the eight groups eight such individuals be selected) and the eight individuals so selected should be randomly assigned to the eight groups. Generally, equal number of items are put in each group so that the size of the group is not likely to affect the result of the study. Variables relating to both populati characteristics are assumed to be randomly distributed among the two groups. Thus, th random replication design is, in fact, an extension of the two-group simple randomiz design.

3.7.5 Randomized Block Design (R.B. Design)

Randomized Block Design is an improvement over the C.R. design. In the R.B. design the princip of local control can be applied along with the other two`principles of experimental designs. In t R.B. design, subjects are first divided into groups, known as blocks, such that within each group t subjects are relatively homogeneous in respect to some selected variable. The variable selected f grouping the subjects is one that is believed to be related to the measures to be obtained in respect the dependent variable. The number of subjects in a given block would be equal to the number treatments and one subject in each block would be randomly assigned to each treatment. In genera blocks are the levels at which we hold the extraneous factor fixed, so that its contribution to the tot variability of data can be measured. The main feature of the R.B. design is that in this treatme appears the same number of times in each block. The R.B. design is analysed by the two-wa analysis of variance technique, discussed later.

Let us illustrate the R.B. design with the help of an example. Suppose four different forms of standardised test in statistics were given to each of five students (selected one from each of the fi I.Q. blocks) and following are the scores which they obtained.

	Very low I.Q.	Low I.Q	Average I.Q.	High I.Q.	Very High I.Q.
	Student A	Student B	Student C	Student D	Student E
Form 1	82	67	57	71	73
Form 2	90	68	54	70	81
Form 3	86	73	51	69	84
Form 4	93	77	60	65	71

Fig. 3.6

If each student separately randomized the order in which he or she took the four tests (by usi random numbers or some similar devices), we refer to the design of this experiment as a R. design. The purpose of this randomization is to take care of such possible extraneous factors (say fatigue) or perhaps the experience gained from repeatedly taking the test.

3.7.6 Latin Square Design (L.S. Design)

This is an experimental design very frequently used in agricultural research. The conditions under which agricultural investigations are carried out are different from those in other studies for nature plays an important role in agriculture. For instance, an experiment has to be made through which the effects of five different varieties of fertilizers on the yield of a certain crop, say wheat, is to be judged. In such a case the varying fertility of the soil in different blocks in which the experiment has to be performed must be taken into consideration; otherwise the results obtained may not be very dependable because the output happens to be the effect of not only fertilizers, but it may also be the effect of fertility of soil. Similarly, there may be impact of varying seeds on the yield. To overcome such difficulties, the L.S. design is used when there are two major extraneous factors such as the varying soil fertility and varying seeds.

The Latin-square design is one wherein each fertilizer, in our example, appears five times but is used only once in each row and in each column of the design. In other words, the treatments in a L.S. design are so allocated among the plots that no treatment occurs more than once in any one row or any one column. The two blocking factors may be represented through rows and columns (one through rows and the other through columns). The following is a diagrammatic form of such a design in respect of, say, five types of fertilizers, viz., A, B, C, D and E and the two blocking factors viz., the varying soil fertility and the varying seeds:

FERTILITY LEVEL

Seeds differences		I	II	III	IV	V
	X_1	A	B	C	D	E
	X_2	B	C	D	E	A
	X_3	C	D	E	A	B
	X_4	D	E	A	B	C
	X_5	E	A	B	C	D

Fig. 3.7

The above diagram clearly shows that in a L.S. design the field is divided into as many blocks as there are varieties of fertilizers and then each block is again divided into as many parts as there are varieties of fertilizers in such a way that each of the fertilizer variety is used in each of the block (whether column-wise or row-wise) only once. The analysis of the L.S. design is very similar to the two-way ANOVA technique.

The merit of this experimental design is that it enables differences in fertility gradients in the field to be eliminated in comparison to the effects of different varieties of fertilizers on the yield of the crop. But this design suffers from one limitation, and it is that although each row and each column represents all fertilizer varieties equally, there may be considerable difference in the row and column means both up and across the field. This, in other words, means that in L.S. design we must assume that there is no interaction between treatments and blocking factors. This defect can, however, b removed by taking the means of rows and columns equal to the field mean by adjusting the results Another limitation of this design is that it requires number of rows, columns and treatments to b equal. This reduces the utility of this design. In case of (2 × 2) L.S. design, there are no degrees o freedom available for the mean square error and hence the design cannot be used. If treatments ar 10 or more, than each row and each column will be larger in size so that rows and columns may no be homogeneous. This may make the application of the principle of local control ineffective. Therefore L.S. design of orders (5 × 5) to (9 × 9) are generally used.

3.7.7 Factorial Designs

Factorial designs are used in experiments where the effects of varying more than one factor are t be determined. They are specially important in several economic and social phenomena where usuall a large number of factors affect a particular problem. Factorial designs can be of two types (i) simple factorial designs and (ii) complex factorial designs. We take them separately.

(i) *Simple factorial designs:* In case of simple factorial designs, we consider the effects o varying two factors on the dependent variable, but when an experiment is done with more than two factors, we use complex factorial designs. Simple factorial design is also terme as a 'two-factor-factorial design', whereas complex factorial design is known as 'multi-factor-factorial design.' Simple factorial design may either be a 2 × 2 simple factoria design, or it may be, say, 3 × 4 or 5 × 3 or the like type of simple factorial design. We illustrate some simple factorial designs as under:

Example 3.1: (2 × 2 simple factorial design). A 2 × 2 simple factorial design can graphically be depicted as follows:

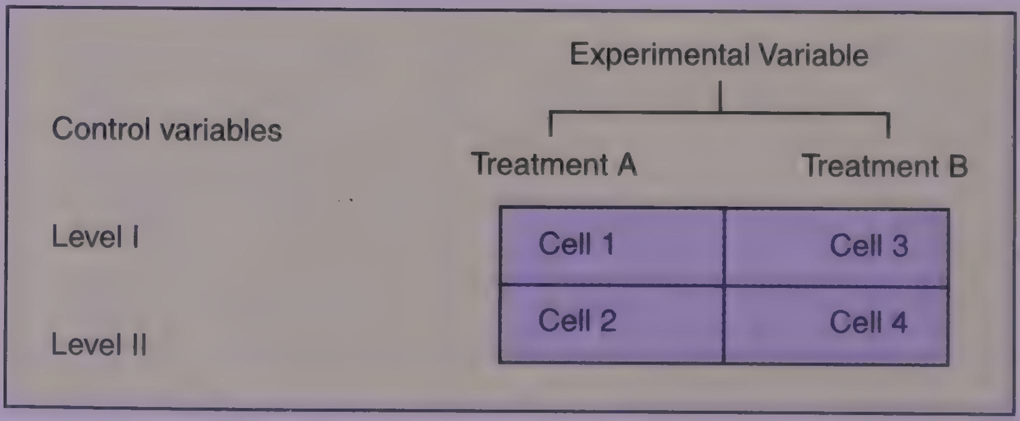


Fig. 3.8: 2 x 2 Simple Factorial Design

In this design the extraneous variable to be controlled by homogeneity is called the control variable and the independent variable, which is manipulated, is called the experimental variable. Then there are two treatments of the experimental variable and two levels of the control variable. As such there are four cells into which the sample is divided. Each of the four combinations would provide one treatment or experimental condition. Subjects are assigned at random to each treatment in the same manner as in a randomized group design. The means for different cells may be obtained along with the means for different rows and columns. Means of different cells represent the mean scores for the dependent variable and the column means in the given design are termed the main effect for treatments without taking into account any differential effect that is due to the level of the control variable. Similarly, the row means in the said design are termed the main effects for levels without regard to treatment. Thus, through this design we can study the main effects of treatments as well as the main effects of levels. An additional merit of this design is that one can examine the interaction between treatments and levels, through which one may say whether the treatment and levels are independent of each other or they are not so. The following examples make clear the interaction effect between treatments and levels. The data obtained in case of two (2 × 2) simple factorial studies may be as given in Fig. 3.9.

STUDY I DATA

		Training		Row Mean
		Treatment A	Treatment B	
Control (Intelligence)	Level I (Low)	15.5	23.3	19.4
	Level II (High)	35.8	30.2	33.0
	Column mean	25.6	26.7	

STUDY II DATA

		Training		Row Mean
		Treatment A	Treatment B	
Control (Intelligence)	Level I (Low)	10.4	20.6	15.5
	Level II (High)	30.6	40.4	35.5
	Column mean	20.5	30.5	

Fig. 3.9

All the above figures (the study I data and the study II data) represent the respective means. Graphically, these can be represented as shown in Fig. 3.10.

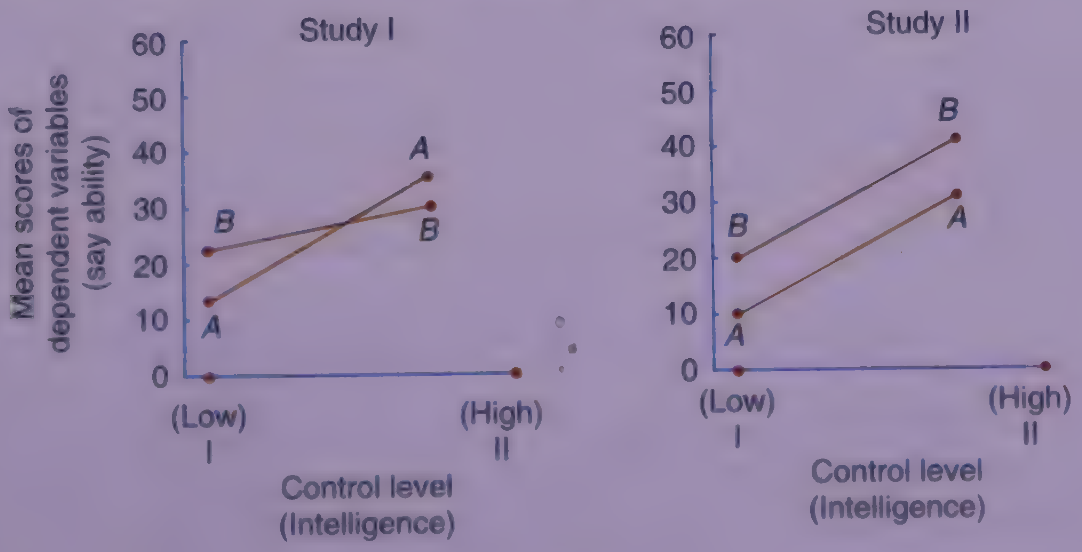


Fig. 3.10

The graph relating to Study I indicates that there is an interaction between the treatment and th level which, in other words, means that the treatment and the level are not independent of each othe The graph relating to Study II shows that there is no interaction effect which means that treatme and level in this study are relatively independent of each other.

The 2 × 2 design need not be restricted in the manner as explained above i.e., having on experimental variable and one control variable, but it may also be of the type having two experiment variables or two control variables. For example, a college teacher compared the effect of the clas size as well as the introduction of the new instruction technique on the learning of research methodolog For this purpose he conducted a study using a 2 × 2 simple factorial design. His design in the graphi form would be as follows:

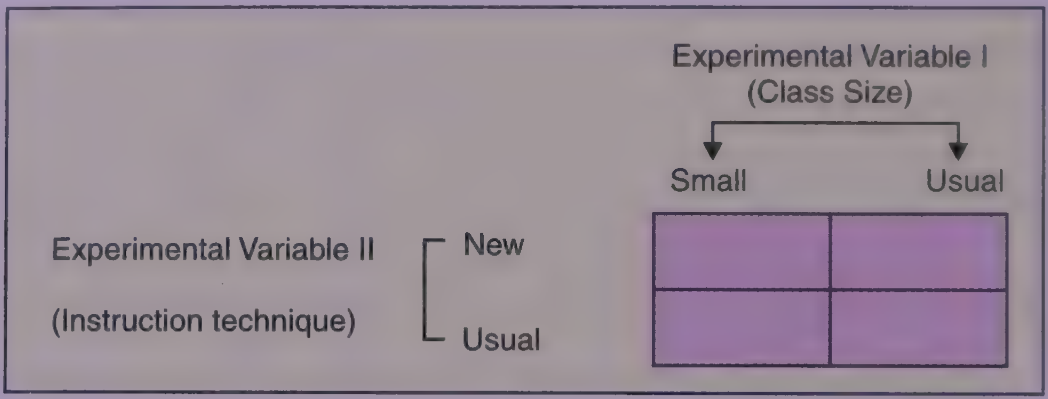


Fig. 3.11

But if the teacher uses a design for comparing male and female and the senior and junior student in the college as they relate to the knowledge of research methodology, in that case we will have 2 × 2 simple factorial design wherein both the variables are control variables as no manipulation i involved in respect of both the variables.

Example 3.2: (4 × 3 simple factorial design). The 4 × 3 simple factorial design will usually includ four treatments of the experimental variable and three levels of the control variable. Graphically may take the following form:

Control Variable	Experimental Variable			
	Treatment A	Treatment B	Treatment C	Treatment D
Level I	Cell 1	Cell 4	Cell 7	Cell 10
Level II	Cell 2	Cell 5	Cell 8	Cell 11
Level III	Cell 3	Cell 6	Cell 9	Cell 12

Fig. 3.12 : **4 x 3 Simple Factorial Design**

This model of a simple factorial design includes four treatments viz., A, B, C, and D of th experimental variable and three levels viz., I, II, and III of the control variable and has 12 differen cells as shown above. This shows that a 2 × 2 simple factorial design can be generalised to an number of treatments and levels. Accordingly, we can name it as such and such (–×–) design. I such a design the means for the columns provide the researcher with an estimate of the main effects for treatments and the means for rows provide an estimate of the main effects for the levels. Such a design also enables the researcher to determine the interaction between treatments and levels.

(ii) *Complex factorial designs:* Experiments with more than two factors at a time involve the use of complex factorial designs. A design which considers three or more independent variables simultaneously is called a complex factorial design. In case of three factors with one experimental variable having two treatments and two control variables, each one of which having two levels, the design used will be termed 2 × 2 × 2 complex factorial design which will contain a total of eight cells as shown below in Fig. 3.13.

2 x 2 x 2 Complex Factorial Design

		Experimental Variable			
		Treatment A		Treatment B	
		Control Variable 2 Level I	Control Variable 2 Level II	Control Variable 2 Level I	Control Variable 2 Level II
Control Variable 1	Level I	Cell 1	Cell 3	Cell 5	Cell 7
	Level II	Cell 2	Cell 4	Cell 6	Cell 8

Fig. 3.13

In Fig. 3.14 a pictorial presentation is given of the design shown below.

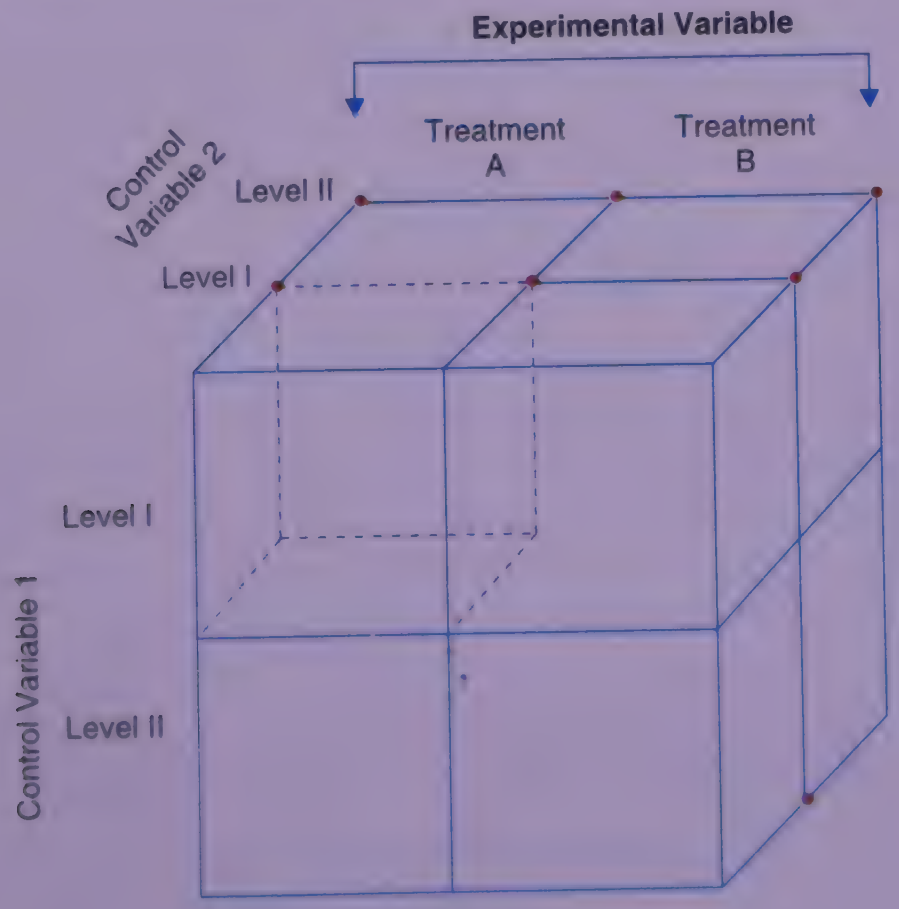


Fig. 3.14

The dotted line cell in the diagram corresponds to Cell 1 of the above stated 2 × 2 × 2 design an is for Treatment A, level I of the control variable 1, and level I of the control variable 2. From thi design it is possible to determine the main effects for three variables i.e., one experimental and tw control variables. The researcher can also determine the interactions between each possible pair o variables (such interactions are called 'First Order interactions') and interaction between variabl taken in triplets (such interactions are called Second Order interactions). In case of a 2 × 2 × design, the further given first order interactions are possible:

Experimental variable with control variable 1 (or EV × CV 1);
Experimental variable with control variable 2 (or EV × CV 2);
Control variable 1 with control variable 2 (or CV1 × CV2);

Three will be one second order interaction as well in the given design (it is between all the thre variables i.e., EV × CV1 × CV2).

To determine the main effects for the experimental variable, the researcher must necessaril compare the combined mean of data in cells 1, 2, 3 and 4 for Treatment A with the combined mea of data in cells 5, 6, 7 and 8 for Treatment B. In this way the main effect for experimental variable independent of control variable 1 and variable 2, is obtained. Similarly, the main effect for contro variable 1, independent of experimental variable and control variable 2, is obtained if we compare th combined mean of data in cells 1, 3, 5 and 7 with the combined mean of data in cells 2, 4, 6 and 8 o our 2 × 2 × 2 factorial design. On similar lines, one can determine the main effect for the contro variable 2 independent of experimental variable and control variable 1, if the combined mean of dat in cells 1, 2, 5 and 6 are compared with the combined mean of data in cells 3, 4, 7 and 8.

To obtain the first order interaction, say, for EV × CV1 in the above stated design, the researche must necessarily ignore control variable 2 for which purpose he may develop 2 × 2 design from the 2 × 2 × 2 design by combining the data of the relevant cells of the latter design as shown in Fig. 3.15

		Experimental Variables	
		Treatment A	Treatment B
Control Variable 1	Level I	Cells 1, 3	Cells 5, 7
	Level II	Cells 2, 4	Cells 6, 8

Fig. 3.15

Similarly, the researcher can determine other first order interactions. The analysis of the first order interaction, in the manner described above, is essentially a sample factorial analysis as only two variables are considered at a time and the remaining one is ignored. But the analysis of the second order interaction would not ignore one of the three independent variables in case of a 2 × 2 × 2 design. The analysis would be termed as a complex factorial analysis.

It may, however, be remembered that the complex factorial design need not necessarily be of 2 × 2 × 2 type design, but can be generalised to any number and combination of experimental and control independent variables. Of course, the greater the number of independent variables included

LIBRARY
NARAYAN RAO MELGIRI
National Law School
BANGALORE

in a complex factorial design, the higher the order of the interaction analysis possible. But the overall task goes on becoming more and more complicated with the inclusion of more and more independent variables in our design.

Factorial designs are used mainly because of the two advantages. (i) They provide equivalent accuracy (as happens in the case of experiments with only one factor) with less labour and as such are a source of economy. Using factorial designs, we can determine the main effects of two (in simple factorial design) or more (in case of complex factorial design) factors (or variables) in one single experiment. (ii) They permit various other comparisons of interest. For example, they give information about such effects which cannot be obtained by treating one single factor at a time. The determination of interaction effects is possible in case of factorial designs.

3.8 CONCLUSION

There are several research designs and the researcher must decide in advance of collection and analysis of data as to which design would prove to be more appropriate for his research project. He must give due weight to various points such as the type of universe and its nature, the objective of his study, the resource list or the sampling frame, desired standard of accuracy and the like when taking a decision in respect of the design for his research project.

PROBLEMS

1. Explain the meaning and significance of a Research design.
2. Explain the meaning of the following in context of Research design.
   (a) Extraneous variables;
   (b) Confounded relationship;
   (c) Research hypothesis;
   (d) Experimental and Control groups;
   (e) Treatments.
3. Describe some of the important research designs used in experimental hypothesis-testing research study.
4. "Research design in exploratory studies must be flexible but in descriptive studies, it must minimise bias and maximise reliability." Discuss.
5. Give your understanding of a good research design. Is single research design suitable in all research studies? If not, why?
6. Explain and illustrate the following research designs:
   (a) Two group simple randomized design;
   (b) Latin square design;
   (c) Random replications design;
   (d) Simple factorial design;
   (e) Informal experimental designs.
7. Write a short note on 'Experience Survey' explaining fully its utility in exploratory research studies.
8. What is research design? Discuss the basis of stratification to be employed in sampling public opinion on inflation.

038805

Design of Sample Surveys

4

Chapter

4.1 INTRODUCTION

All items in any field of inquiry constitute a 'Universe' or 'Population.' A complete enumeration of all items in the 'population' is known as a census inquiry or census survey. It is obvious that for any study or investigation census survey is rather infeasible. For example, to have an idea of average per capita monthly income of the people in India, we will have to enumerate all the earning individuals in the country, which is a very difficult task.

Census survey is impossible in the situations when population is infinite. In some cases when population is finite but the units are destroyed while inspected, census survey is not at all desirable, e.g., inspection of crackers. Further, many a time it is not possible to examine every item in the population, and sometimes it is possible to obtain sufficiently accurate results by studying only a part of total population. In such cases there is no utility of census surveys.

However, it needs to be emphasised that when the universe is a small one, it is no use resorting to a sample survey. When field studies are undertaken in practical life, considerations of time and cost almost invariably lead to a selection of respondents i.e., selection of only a few items. The respondents selected should be as representative of the total population as possible in order to produce a miniature cross-section. The selected respondents constitute what is technically called a 'sample' and the selection process is called 'sampling technique.' The survey so conducted is known as 'sample survey'. Algebraically, let the population size be N and if a part of size n (which is $< N$) of this population is selected according to some rule for studying some characteristic of the population, the group consisting of these n units is known as 'sample'. Researcher must prepare a sample design for his study i.e., he must plan how a sample should be selected and of what size such a sample would be.

4.2 SAMPLE DESIGN

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of items to be included in the sample i.e., the size of the sample. Sample design is determined before data are collected. There are many sample designs from which a researcher can choose. Some designs are relatively more precise and easier to apply than others. Researcher must select/prepare a sample design which should be reliable and appropriate for his research study. The main steps of sampling design are as follows:

(i) **Objective:** The first step of sampling design is to define the objectives of survey in clear and concrete terms. The sponsors or the researchers of the survey should confirm that the objectives are commensurate with the money, manpower and time limit available for the survey.

(ii) **Population:** In order to meet the objectives of the survey, what should be the population? This question should be answered in the second step. The population should be clearly defined.

(iii) **Sampling units and frame:** A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of such units that he has to select for his study. The list of sampling units is called as 'frame' or sampling frame. Sampling frame contains the names of all items of a universe (in case of finite universe only).

Such a list should be comprehensive, correct, reliable and appropriate. It is extremely important for the source list to be as representative of the population as possible.

(iv) **Size of sample:** This refers to the number of items to be selected from the universe to constitute a sample. This is a major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate. The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed. The size of population must be kept in view for this also limits the sample size. The parameters of interest in a research study must be kept in view, while deciding the size of the sample. Costs too dictate the size of sample that we can draw. As such, budgetary constraint must invariably be taken into consideration when we decide the sample size. The formulation of sample size determination is provided in Chapter 10.

(v) **Parameters of interest:** Statistical constants of the population are called as parameters, e.g., population mean, population proportion etc. When we do census survey we get the actual value of parameters. On the other hand, when we do sample survey we get the estimates of unknown population parameters in place of their actual values.

In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population. There may also be important sub-groups in the population about whom we would like to make estimates. All this has a strong impact upon the sample design we would accept.

(vi) **Data collection:** No irrelevant information should be collected and no essential informatio should be discarded. The objectives of the survey should be very much clear in the mind c surveyor. The methods of data collection are discussed in Chapter 6.

(vii) **Non-respondents:** Because of practical difficulties, data may not be collected for all th sampled units. This non-response tends to change the results. The reasons for non-respons should be recorded by the investigator. Such cases should be handled with caution.

(viii) **Selection of proper sampling design:** The researcher must decide the type of sampl he will use i.e., he must decide about the technique to be used in selecting the items for th sample. In fact, this technique or procedure stands for the sample design itself. There ar several sample designs (explained in the pages that follow) out of which the researche must choose one for his study. Obviously, he must select that design which, for a give sample size and for a given cost, has a smaller sampling error.

(ix) **Organizing field work:** The success of a survey depends on the reliable field work There should be efficient supervisory staff and trained personnels for the field work.

(x) **Pilot survey:** It is always helpful to try out the research design on a small scale before going to the field. This is called as 'pilot survey' or 'pretest'. It might give the better idea o practical problems and troubles.

(xi) **Budgetary constraint:** Cost considerations, from practical point of view, have a majo impact upon decisions relating to not only the size of the sample but also to the type o sample. This fact can even lead to the use of a non-probability sample.

4.3 SAMPLING AND NON-SAMPLING ERRORS

The errors involved in the collection of data are classified into sampling and non-sampling errors.

4.3.1 Sampling Errors

Sampling errors arise due to the fact that only a part of the population has been used to estimate population parameters and to draw inferences about the population. Sampling errors are absent in a census survey. *Sampling error* can be measured for a given sample design and size. The measurement of sampling error is usually called the 'precision of the sampling plan'. If we increase the sample size, the precision can be improved. But increasing the size of the sample has its own limitations viz., a large sized sample increases the cost of collecting data and also enhances the systematic bias. Thus the effective way to increase precision is usually to select a better sampling design which has a smaller sampling error for a given sample size at a given cost. In practice, however, people prefer a less precise design because it is easier to adopt the same and also because of the fact that systematic bias can be controlled in a better way in such a design.

In brief, *while selecting a sampling procedure, researcher must ensure that the procedure causes a relatively small sampling error and helps to control the systematic bias in a better way.*

4.3.2 Non-sampling Errors

Non-sampling errors arise at the stage of collection and preparation of data and thus are present in both the sample survey as well as the census survey. Thus the data obtained in census survey is free from sampling errors, however subjected to non-sampling errors. Non-sampling errors can be reduced by defining the sampling units, frame and the population correctly and by employing efficient people in the investigations.

4.4 SAMPLE SURVEY VS. CENSUS SURVEY

In a sample survey, since we study only a subpart of the whole population, requires less money and less time. Most of the times, non-sampling errors are so much large that the results of sample survey are much more accurate than those of census survey. Non-sampling errors arise due to a number of factors such as inefficiency of field workers, non-response, bias due to interviewers, etc. These errors are likely to grow when the number of units inspected increases.

However, if the objective of study is very serious in nature and information is required about each and every sampling unit, there is no way out but to resort to census survey. Moreover, if time and money are not important factors or if population is not so large, a census survey may provide better results than any sample survey, provided efficient and trained staff is employed.

4.5 TYPES OF SAMPLING DESIGNS

The method of selecting a sample is of fundamental importance and depends upon the nature of data and investigation. The techniques of selecting a sample are classified as – ‘non - probability sampling’ and ‘probability sampling,’. We take up these two designs separately.

4.5.1 Non-probability Sampling

Non-probability sampling is that sampling procedure which does not afford any basis for estimating the probability that each item in the population has of being included in the sample. Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgement sampling. In this type of sampling, items for the sample are selected deliberately by the researcher; his choice concerning the items remain supreme. In other words, under non-probability sampling the organisers of the inquiry purposively choose the particular units of the universe for constituting a sample on the basis that the small mass that they so select out of a huge one will be typical or representative of the whole. For instance, if economic conditions of people living in a state are to be studied, a few towns and villages may be purposively selected for intensive study on the principle that they can be representative of the entire state. Thus, the judgement of the organisers of the study plays an important part in this sampling design.

In such a design, personal element has a great chance of entering into the selection of the sample. The investigator may select a sample which shall yield results favourable to his point of view and if that happens, the entire inquiry may get vitiated. Thus, there is always the danger of bias entering into this type of sampling technique. But in the investigators are impartial, work without bias and have the necessary experience so as to take sound judgement, the results obtained from a analysis of deliberately selected sample may be tolerably reliable. However, in such a sampling, ther is no assurance that every element has some specifiable chance of being included. Sampling error i this type of sampling cannot be estimated and the element of bias, great or small, is always there. A such this sampling design is gets rarely adopted in large inquiries of importance. However, in sma inquiries and researches by individuals, this design may be adopted because of the relative advantag of time and money inherent in this method of sampling. *Quota sampling* is also an example of non probability sampling. Under quota sampling the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled. In other words, the actua selection of the items for the sample is left to the interviewer's discretion. This type of sampling i very convenient and is relatively inexpensive. But the samples so selected certainly do not posses the characteristic of random samples. Quota samples are essentially judgement samples and inference drawn on their basis are not amenable to statistical treatment in a formal way.

4.5.2 Probability Sampling

Probability sampling is also known as 'random sampling' or 'chance sampling'. Under this samplin design, every item of the universe has an equal chance of inclusion in the sample. It is, so to say, lottery method in which individual units are picked up from the whole group not deliberately but b some mechanical process. Here it is blind chance alone that determines whether one item or th other is selected. The results obtained from probability or random sampling can be assured in term of probability i.e., we can measure the errors of estimation or the significance of results obtaine from a random sample, and this fact brings out the superiority of random sampling design over th deliberate sampling design. Random sampling ensures the law of Statistical Regularity which state that if on an average the sample chosen is a random one, the sample will have the same compositio and characteristics as the universe. This is the reason why random sampling is considered as the bes technique of selecting a representative sample.

4.5.2.1 Simple Random Sampling

Simple random sampling from a finite population refers to that method of sample selection which gives each possible sample combination an equal probability of being picked up and each item in the entire population to have an equal chance of being included in the sample. This applies to sampling without replacement i.e., once an item is selected for the sample, it cannot appear in the sample again (sampling with replacement is used frequently less in which procedure the element selected for the sample is returned to the population before the next element is selected. In such a situation the same element could appear twice in the same sample before the second element is chosen). In brief, the implications of random sampling (or simple random sampling) are:

(a) It gives each element in the population an equal probability of getting into the sample; and all choices are independent of one another.

(b) It gives each possible sample combination an equal probability of being chosen.

Keeping this in view we can define a simple random sample (or simply a random sample) from a finite population as a sample which is chosen in such a way that each of the ${}^{N}C_{n}$ possible samples has the same probability, $1/{}^{N}C_{n}$, of being selected. To make it more clear we take a certain finite opulation consisting of six elements (say *a*, *b*, *c*, *d*, *e*, *f*) i.e., $N = 6$. Suppose that we want to take a ample of size $n = 3$ from it. Then there are $^6C_3 = 20$ possible distinct samples of the required size, nd they consist of the elements *abc*, *abd*, *abe*, *abf*, *acd*, *ace*, *acf*, *ade*, *adf*, *aef*, *bcd*, *bce*, *bcf*, *bde*, *df*, *bef*, *cde*, *cdf*, *cef*, and *def*. If we choose one of these samples in such a way that each has the robability 1/20 of being chosen, we will then call this a random sample or a simple random sample.

With regard to the question of how to take a random sample in actual practice, we could, in imple cases like the one above, write each of the possible samples on a slip of paper, mix these slips horoughly in a container and then draw as a lottery either blindfolded or by rotating a drum or by any ther similar device. Such a procedure is obviously impractical, if not altogether impossible in complex roblems of sampling. In fact, the practical utility of such a method is very much limited.

Fortunately, we can take a random sample in a relatively easier way without taking the trouble of nlisting all possible samples on paper-slips as explained above. Instead of this, we can write the ame of each element of a finite population on a slip of paper, put the slips of paper so prepared into box or a bag and mix them thoroughly and then draw (without looking) the required number of slips or the sample one after the other without replacement. In doing so we must make sure that in successive drawings each of the remaining elements of the population has the same chance of being selected. This procedure will also result in the same probability for each possible sample. We can verify this by taking the above example. Since we have a finite population of 6 elements and we want to select a sample of size 3, the probability of drawing any one element for our sample in the first draw is 3/6, the probability of drawing one more element in the second draw is 2/5 (the first element drawn is not replaced) and similarly the probability of drawing one more element in the third draw is 1/4. Since these draws are independent, the joint probability of the three elements which constitute our sample is the product of their individual probabilities and this works out to $3/6 \times 2/5 \times 1/4 = 1/20$. This verifies our earlier calculation.

Even this relatively easy method of obtaining a random sample can be simplified in actual practice by the use of random number tables. Various statisticians like Tippett, Yates, Fisher have prepared tables of random numbers which can be used for selecting a random sample. Generally, Tippett's random number tables are used for the purpose. Tippett gave10,400 four figure numbers. He selected 41,600 digits from the census reports and combined them into fours to give his random numbers which may be used to obtain a random sample.

We can illustrate the procedure by an example. First of all we reproduce the first thirty sets of Tippett's numbers

2952	6641	3992	9792	7979	5911
3170	5624	4167	9525	1545	1396
7203	5356	1300	2693	2370	7483
3408	2769	3563	6107	6913	7691
0560	5246	1112	9025	6008	8126

Suppose we are interested in taking a sample of 10 units from a population of 5000 units, bearing numbers from 3001 to 8000. We shall select 10 such figures from the above random numbers which are not less than 3001 and not greater than 8000. If we randomly decide to read the table numbers from left to right, starting from the first row itself, we obtain the following numbers: 6641, 3992, 7979, 5911, 3170, 5624, 4167, 7203, 5356, and 7483.

The units bearing the above serial numbers would then constitute our required random sampl

One may note that it is easy to draw random samples from finite populations with the aid random number tables only when lists are available and items are readily numbered. But in son situations it is often impossible to proceed in the way we have narrated above. For example, if w want to estimate the mean height of trees in a forest, it would not be possible to number the trees, an choose random numbers to select a random sample. In such situations what we should do is to sele some trees for the sample haphazardly without aim or purpose, and should treat the sample as random sample for study purposes.

Other alternative may be to divide the whole forest into a finite number of equal-sized blocks an consider these blocks as sampling units. Height of each and every tree falling into selected block should be measured.

Simple random sample, we discussed above in 'simple random sample without replacement Here, a unit which is selected once is not replaced back to the population, i.e., one sampling unit ca be selected only once. Other alternative to this design is 'simple random sample with replacement' Because of the possibility of over-representation of some sampling units, this design is not of muc practical usage.

So far we have talked about random sampling, keeping in view only the finite populations. Bu what about random sampling in context of infinite populations? It is relatively difficult to explain th concept of random sampling from an infinite population. However, a few examples will show th basic characteristic of such a sample. Suppose we consider the 20 throws of a fair dice as a sampl from the hypothetically infinite population which consists of the results of all possible throws of th dice. If the probability of getting a particular number, say 1, is the same for each throw and the 2 throws are all independent, then we say that the sample is random. Similarly, it would be said to b sampling from an infinite population if we sample with replacement from a finite population and ou sample would be considered as a random sample if in each draw all elements of the population hav the same probability of being selected and successive draws happen to be independent. In brief, on can say that the selection of each item in a random sample from an infinite population is controlled b the same probabilities and that successive selections are independent of one another.

4.5.3 Complex Random Sampling Designs

Some complex random sampling designs, which are the mixture of probability and non-probabilit sampling methods, are given below.

4.5.3.1 Systematic Sampling

In some instances, the most practical way of sampling is to select every *i*th item on a list. Sampling of this type is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random numbers to pick up the unit with which to start. For instance, if a 4 per cent sample is desired, the first item would be selected randomly from the first twenty-five and thereafter every 25th item would automatically be included in the sample. Thus, in systematic sampling only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals. Although a systematic sample is not a random sample in the strict sense of the term, but it is often considered reasonable to treat systematic sample as if it were a random sample.

Systematic sampling has certain plus points. It can be taken as an improvement over a simple ndom sample in as much as the systematic sample is spread more evenly over the entire population. is an easier and less costlier method of sampling and can be conveniently used even in case of rge populations. But there are certain dangers too in using this type of sampling. If there is a hidden eriodicity in the population, systematic sampling will prove to be an inefficient method of sampling. or instance, every 25th item produced by a certain production process is defective. If we are to elect a 4% sample of the items of this process in a systematic manner, we would either get all efective items or all good items in our sample depending upon the random starting position. If all ements of the universe are ordered in a manner representative of the total population, i.e., the opulation list is in random order, systematic sampling is considered equivalent to random sampling. ut if this is not so, then the results of such sampling may, at times, not be very reliable. In practice, ystematic sampling is used when lists of population are available and they are of considerable ength.

.5.3.2 *Stratified Sampling*

a population from which a sample is to be drawn does not constitute a homogeneous group, tratified sampling technique is generally applied in order to obtain a representative sample. Under tratified sampling the population is divided into several sub-populations that are individually more omogeneous than the total population (the different sub-populations are called 'strata') and then we elect items from each stratum to constitute a sample. Since each stratum is more homogeneous han the total population, we are able to get more precise estimates for each stratum and by estimating nore accurately each of the component parts, we get a better estimate of the whole. In brief, tratified sampling results in more reliable and detailed information.

The following three questions are highly relevant in the context of stratified sampling:

(a) How to form strata?

(b) How should items be selected from each stratum?

(c) How many items be selected from each stratum or how to allocate the sample size of each stratum?

Regarding the first question, we can say that the strata be formed on the basis of common haracteristic(s) of the items to be put in each stratum. This means that various strata be formed in uch a way as to ensure elements being most homogeneous within each stratum and most eterogeneous between the different strata. Thus, strata are purposively formed and are usually ased on past experience and personal judgement of the researcher. One should always remember hat careful consideration of the relationship between the characteristics of the population and the haracteristics to be estimated are normally used to define the strata. At times, pilot study may be onducted for determining a more appropriate and efficient stratification plan. We can do so by aking small samples of equal size from each of the proposed strata and then examining the variances vithin and among the possible stratifications, we can decide an appropriate stratification plan for our nquiry.

In respect of the second question, we can say that the usual method, for selection of items for the ample from each stratum, resorted to is that of simple random sampling. Systematic sampling can e used if it is considered more appropriate in certain situations.

Regarding the third question, we usually follow the method of proportional allocation under whi the sizes of the samples from the different strata are kept proportional to the sizes of the strata. Th is, if P_i represents the proportion of population included in stratum i, and n represents the total samp size, the number of elements selected from stratum i is $n \,.\, P_i$. To illustrate it, let us suppose that v want a sample of size $n = 30$ to be drawn from a population of size $N = 8000$ which is divided in three strata of size $N_1 = 4000$, $N_2 = 2400$ and $N_3 = 1600$. Adopting proportional allocation, we shall g the sample sizes as under for the different strata:

For strata with $N_1 = 4000$, we have $P_1 = 4000/8000$

and hence $n_1 = n \,.\, P_1 = 30\ (4000/8000) = 15$

Similarly, for strata with $N_2 = 2400$, we have

$n_2 = n \,.\, P_2 = 30\ (2400/8000) = 9$, and

for strata with $N_3 = 1600$, we have

$n_3 = n \,.\, P_3 = 30\ (1600/8000) = 6$.

Thus, using proportional allocation, the sample sizes for different strata are 15, 9 and 6 respectivel which is in proportion to the sizes of the strata viz., 4000 : 2400 : 1600. Proportional allocation considered most efficient and an optimal design when the cost of selecting an item is equal for eac stratum, there is no difference in within-stratum variances, and the purpose of sampling happens t estimate the population value of some characteristic. But in case the purpose happens to compar the differences among the strata, then equal sample selection from each stratum would be mor efficient even if the strata differ in sizes. In cases where strata differ not only in size but also i variability and it is considered reasonable to take larger samples from the more variable strata an smaller samples from the less variable strata, we can then account for both (differences in stratu size and differences in stratum variability) by using disproportionate sampling design by requiring:

$$n_1/N_1\sigma_1 = n_2/N_2\sigma_2 = \ldots\ldots\ldots = n_k/N_k\sigma_k$$

where $\sigma_1, \sigma_2, \ldots$ and σ_k denote the standard deviations of the k strata, $N_1, N_2, \ldots, N_k$ denote th sizes of the k strata and $n_1, n_2, \ldots, n_k$ denote the sample sizes of k strata. This is called '*optimu allocation*' in the context of disproportionate sampling. The allocation in such a situation results i the following formula for determining the sample sizes different strata:

$$n_i = \frac{n \cdot N_i\,\sigma_i}{N_1\,\sigma_1 + N_2\,\sigma_2 + \ldots + N_k\,\sigma_k} \qquad \text{for } i = 1, 2, \ldots \text{ and } k.$$

We may illustrate the use of this by an example.

Example 4.1: A population is divided into three strata so that $N_1 = 5000$, $N_2 = 2000$ an $N_3 = 3000$. Respective standard deviations are:

$$\sigma_1 = 15,\ \sigma_2 = 18 \text{ and } \sigma_3 = 5.$$

How should a sample of size $n = 84$ be allocated to the three strata, if we want optimum allocatio using disproportionate sampling design?

Solution: Using the disproportionate sampling design for optimum allocation, the sample sizes fo different strata will be determined as under:

Sample size for strata with $N_1 = 5000$

$$n_1 = \frac{84(5000)\,(15)}{(5000)\,(15) + (2000)\,(18) + (3000)\,(5)}$$

$$= 6300000/126000 = 50$$

Sample size for strata with $N_2 = 2000$

$$n_2 = \frac{84(2000)\,(18)}{(5000)\,(15) + (2000)\,(18) + (3000)\,(5)}$$

$$= 3024000/126000 = 24$$

Sample size for strata with $N_3 = 3000$

$$n_3 = \frac{84(3000)\,(5)}{(5000)\,(15) + (2000)\,(18) + (3000)\,(5)}$$

$$= 1260000/126000 = 10$$

In addition to differences in stratum size and differences in stratum variability, we may have ifferences in stratum sampling cost, then we can have cost optimal disproportionate sampling design y requiring

$$\frac{n_1}{N_1\,\sigma_1\sqrt{C_1}} = \frac{n_2}{N_2\,\sigma_2\sqrt{C_2}} = \ldots = \frac{n_k}{N_k\,\sigma_k\sqrt{C_k}}$$

vhere

C_1 = Cost of sampling in stratum 1

C_2 = Cost of sampling in stratum 2

⋮

C_k = Cost of sampling in stratum k

nd all other terms remain the same as explained earlier. The allocation in such a situation results in he following formula for determining the sample sizes for different strata:

$$n_i = \frac{n \cdot N_i\sigma_i/\sqrt{C_i}}{N_1\sigma_1/\sqrt{C_1} + N_2\sigma_2/\sqrt{C_2} + \ldots + N_k\sigma_k/\sqrt{C_k}} \quad \text{for } i = 1, 2, \ldots, k$$

It is not necessary that stratification be done keeping in view a single characteristic. Populations re often stratified according to several characteristics. For example, a system-wide survey designed o determine the attitude of students towards a new teaching plan, a state college system with 20 colleges might stratify the students with respect to class, section and college. Stratification of this ype is known as *cross-stratification*, and up to a point such stratification increases the reliability of estimates and is much used in opinion surveys.

From what has been stated above in respect of stratified sampling, we can say that the sample so constituted is the result of successive application of purposive (involved in stratification of items) and andom sampling methods. As such it is an example of mixed sampling. The procedure wherein we irst have stratification and then simple random sampling is known as stratified random sampling.

4.5.3.3 Cluster Sampling

If the total area of interest happens to be a big one, a convenient way in which a sample can be tak is to divide the area into a number of smaller non-overlapping areas and then to randomly selec number of these smaller areas (usually called clusters), with the ultimate sample consisting of all samples of) units in these small areas or clusters.

Thus, in cluster sampling the total population is divided into a number of relatively small su divisions which are themselves clusters of still smaller units and then some of these clusters randomly selected for inclusion in the overall sample. Suppose, we want to estimate the proportion machine-parts in an inventory which are defective. Also assume that there are 20,000 machine pa in the inventory at a given point of time, stored in 400 cases of 50 each. Now, using a clus sampling, we would consider the 400 cases as clusters and randomly select '*n*' cases and examine the machine-parts in each randomly selected case.

Cluster sampling, no doubt, reduces cost by concentrating surveys in selected clusters. B certainly it is less precise than random sampling. There is also not as much information in '*n*' observatio within a cluster as there happens to be in '*n*' randomly drawn observations. Cluster sampling is us only because of the economic advantage it possesses; estimates based on cluster samples are usual more reliable per unit cost.

If clusters happen to be some geographic sub-divisions, in that case cluster sampling is bett known as area sampling. In other words, cluster designs, where the primary sampling unit represen a cluster of units based on geographic area, are distinguished as area sampling. The plus and min points of cluster sampling are also applicable to area sampling.

4.5.3.4 Multi-stage Sampling

Multi-stage sampling is a further development of the principle of cluster sampling. Suppose, we wa to investigate the working efficiency of nationalised banks in India and we want to take a sample few banks for this purpose. The first stage is to select large primary sampling unit such as states a country. Then we may select certain districts and interview all banks in the chosen districts. Th would represent a two-stage sampling design with the ultimate sampling units being clusters districts.

If instead of taking a census of all banks within the selected districts, we select certain towns an interview all banks in the chosen towns. This would represent a three-stage sampling design. instead of taking a census of all banks within the selected towns, we randomly sample banks fro each selected town, then it is a case of using a four-stage sampling plan. If we select randomly at a stages, we will have what is known as 'multistage random sampling design'.

Ordinarily multi-stage sampling is applied in big inquires extending to a considerable larg geographical area, say, the entire country. There are two advantages of this sampling design viz (a) It is easier to administer than most single stage designs mainly because of the fact that samplin frame under multi-stage sampling is developed in partial units. (b) A large number of units can b sampled for a given cost under multi-stage sampling because of sequential clustering, whereas this i not possible in most of the simple designs.

5.3.5 Sampling with Probability Proportional to Size

case the cluster sampling units do not have the same number or approximately the same number elements, it is considered appropriate to use a random selection process where the probability of ch cluster being included in the sample is proportional to the size of the cluster. For this purpose, e have to list the number of elements in each cluster irrespective of the method of ordering the uster. Then we must sample systematically the appropriate number of elements from the cumulative tals. The actual numbers selected in this way do not refer to individual elements, but indicate which usters and how many from the cluster are to be selected by simple random sampling or by systematic mpling. The results of this type of sampling are equivalent to those of a simple random sample and e method is less cumbersome and is also relatively less expensive. We can illustrate this with the lp of an example.

cample 4.2: The following are the number of departmental stores in 15 cities: 35, 17, 10, 32, 70, , 26, 19, 26, 66, 37, 44, 33, 29 and 28. If we want to select a sample of 10 stores, using cities as usters and selecting within clusters proportional to size, how many stores from each city should be osen? (Use a starting point of 10).

lution: Let us put the information as under (Table 4.1):

Since in the given problem, we have 500 departmental stores from which we have to select a mple of 10 stores, the appropriate sampling interval is 50. As we have to use the starting point of 0*, so we add successively increments of 50 till 10 numbers have been selected. The numbers, thus, tained are: 10, 60, 110, 160, 210, 260, 310, 360, 410 and 460 which have been shown in the last lumn of the table (Table 4.1) against the concerning cumulative totals. From this we can say that o stores should be selected randomly from city number five and one each from city number 1, 3, 7, 10, 11, 12, and 14. This sample of 10 stores is the sample with probability proportional to size.

Table 4.1

City number	*No. of departmental stores*	*Cumulative total*	*Sample*
1	35	35	10
2	17	52	
3	10	62	60
4	32	94	
5	70	164	110
6	28	192	160
7	26	218	210
8	19	237	
9	26	263	260
10	66	329	310
11	37	366	360
12	44	410	410
13	33	443	
14	29	472	460
15	28	500	

---

*If the starting point is not mentioned, then the same can randomly be selected.

4.5.3.6 Sequential Sampling

This sampling design is some what complex sample design. The ultimate size of the sample under t technique is not fixed in advance, but is determined according to mathematical decision rules on basis of information yielded as survey progresses. This is usually adopted in case of acceptan sampling plan in context of statistical quality control. When a particular lot is to be accepted rejected on the basis of a single sample, it is known as single sampling; when the decision is to taken on the basis of two samples, it is known as double sampling and in case the decision rests the basis of more than two samples but the number of samples is certain and decided in advance, sampling is known as multiple sampling. But when the number of samples is more than two but i neither certain nor decided in advance, this type of system is often referred to as sequential sampli Thus, in brief, we can say that in sequential sampling, one can go on taking samples one after anot as long as one desires to do so.

REMARK

From a brief description of the various sample designs presented above, we can say that normall one should resort to simple random sampling because under it bias is generally eliminated and th sampling error can be estimated. But purposive sampling is considered more appropriate when th universe happens to be small and a known characteristic of it is to be studied intensively. Ther are situations in real life under which sample designs other than simple random samples may b considered better (say easier to obtain, cheaper or more informative) and as such the same ma be used. In a situation when random sampling is not possible, then we have to use necessarily sampling design other than random sampling. At times, several methods of sampling may well b used in the same study.

PROBLEMS

1. What do you mean by 'Sample Design'? What points should be taken into consideration by a researche in developing a sample design for this research project.
2. How would you differentiate between simple random sampling and complex random sampling designs Explain clearly giving examples.
3. Why probability sampling is generally preferred in comparison to non-probability sampling? Explai the procedure of selecting a simple random sample.
4. Under what circumstances stratified random sampling design is considered appropriate? How woul you select such sample? Explain by means of an example.
5. Distinguish between:
   - (a) Restricted and unrestricted sampling;
   - (b) Convenience and purposive sampling;
   - (c) Systematic and stratified sampling;
   - (d) Cluster and area sampling.

. Under what circumstances would you recommend:
  (a) A probability sample?
  (b) A non-probability sample?
  (c) A stratified sample?
  (d) A cluster sample?

7. Explain and illustrate the procedure of selecting a random sample.

. How will you optimize sampling and non-sampling errors?

9. "A systematic bias results from errors in the sampling procedures". What do you mean by such a systematic bias? Describe the important causes responsible for such a bias.

0. (a) The following are the number of departmental stores in 10 cities: 35, 27, 24, 32, 42, 30, 34, 40, 29 and 38. If we want to select a sample of 15 stores using cities as clusters and selecting within clusters proportional to size, how many stores from each city should be chosen? (Use a starting point of 4).
  (b) What sampling design might be used to estimate the weight of a group of men and women?

1. A certain population is divided into five strata so that $N_1 = 2000$, $N_2 = 2000$, $N_3 = 1800$, $N_4 = 1700$, and $N_5 = 2500$. Respective standard deviations are: $\sigma_1 = 1.6$, $\sigma_2 = 2.0$, $\sigma_3 = 4.4$, $\sigma_4 = 4.8$, $\sigma_5 = 6.0$ and further the expected sampling cost in the first two strata is Rs 4 per interview and in the remaining three strata the sampling cost is Rs 6 per interview. How should a sample of size $n = 226$ be allocated to five strata if we adopt proportionate sampling design; if we adopt disproportionate sampling design considering (i) only the differences in stratum variability (ii) differences in stratum variability as well as the differences in stratum sampling costs.

5

Chapte

Measurement and Scaling

In our daily life as we are said to measure when we use some yardstick to determine weight, hei or some other features of a physical object. We also measure when we judge how well we lik song, a painting or the personalities of our friends. We, thus, measure physical objects as well abstract concepts. Measurement is a relatively complex and demanding task, specially so whe concerns qualitative or abstract phenomena. Other examples of qualitative characteristics are ta honesty, intelligence, customer's perception and brand loyalty etc. These characteristics are a called as constructs. The feelings and perceptions of customers and employees are very import for a company.

It is easy to assign numbers in respect of properties of some objects, but it is relatively difficul respect of others. For instance, measuring such things as social conformity, intelligence, or mar adjustment is much less obvious and requires much closer attention than measuring physical weig biological age or a person's financial assets. In other words, properties like weight, height, etc., c be measured directly with some standard unit of measurement, but it is not that easy to meas properties like motivation to succeed, ability to stand stress and the like. For the meaningful assessm of the qualitative characteristics it is essential that they are also measured.

5.1 QUANTITATIVE AND QUALITATIVE DATA

Measurement is defined as a process of associating numbers or symbols to observations obtained a research study. These observations could be qualitative or quantitative.

Most of the analysis can be conducted using quantitative data. For example, mean, standa deviation, etc. can be computed for quantitative characteristics. Qualitative characteristics can counted and cannot be computed. Therefore, the researcher must have a clear understanding of t type of characteristic or variable before collecting the data. The observations on qualitative variabl may also be assigned numbers. For example, we can record a person's marital status as 1, 2, 3 o depending on whether the person is single, married, widowed or divorced. We can as well rec "Yes or No" answers to a question as "0" and "1" (or as 1 and 2 or perhaps as 59 and 60). In t artificial or nominal way, categorical data (qualitative or descriptive) can be made into numerical d and if we thus code the various categories, we refer to the numbers we record as nominal da *minal data* is numerical in name only, because they do not share any of the properties of the nbers we deal in ordinary arithmetic. For instance if we record marital status as 1, 2, 3, or 4 as ed above, we cannot write $4 > 2$ or $3 < 4$ and we cannot write $3 - 1 = 4 - 2$, $1 + 3 = 4$ or $4 \div 2 = 2$.

In those situations when we cannot do anything except set up inequalities, we refer to the data as *linal data.* For instance, if one mineral can scratch another, it receives a higher hardness number l on Mohs' scale the numbers from 1 to 10 are assigned respectively to talc, gypsum, calcite, orite, apatite, feldspar, quartz, topaz, sapphire and diamond. With these numbers we can write 2 or $6 < 9$ as apatite is harder than gypsum and feldspar is softer than sapphire, but we cannot te for example $10 - 9 = 5 - 4$, because the difference in hardness between diamond and sapphire ctually much greater than that between apatite and fluorite. It would also be meaningless to say t topaz is twice as hard as fluorite simply because their respective hardness numbers on Mohs' le are 8 and 4. The greater than symbol (i.e., >) in connection with ordinal data may be used to signate "happier than" "preferred to" and so on.

When in addition to setting up inequalities we can also form differences, we refer to the data as *erval data.* Suppose we are given the following temperature readings (in degrees Fahrenheit): °, 63°, 70°, 95°, 110°, 126° and 135°. In this case, we can write $100^\circ > 70^\circ$ or $95^\circ < 135^\circ$ which nply means that 110° is warmer than 70° and that 95° is cooler than 135°. We can also write for ample $95^\circ - 70^\circ = 135^\circ - 110^\circ$, since equal temperature differences are equal in the sense that the me amount of heat is required to raise the temperature of an object from 70° to 95° or from 110° 135°. On the other hand, it would not mean much if we said that 126° is twice as hot as 63°, even ough $126^\circ \div 63^\circ = 2$. To show the reason, we have only to change to the centigrade scale, where first temperature becomes $5/9\ (126 - 32) = 52^\circ$, the second temperature becomes 5/9 $3 - 32) = 17^\circ$ and the first figure is now more than three times the second. This difficulty arises m the fact that Fahrenheit and Centigrade scales both have artificial origins (zeros) i.e., the mber 0 of neither scale is indicative of the absence of whatever quantity we are trying to measure.

When in addition to setting up inequalities and forming differences we can also form quotients ., when we can perform all the customary operations of mathematics), we refer to such data as *tio data.* In this sense, ratio data includes all the usual measurement (or determinations) of length, ight, money amounts, weight, volume, area, pressures etc.

The above stated distinction between nominal, ordinal, interval and ratio data is important for the ture of a set of data may suggest the use of particular statistical techniques[*]. A researcher has to quite alert about this aspect while measuring properties of objects or of abstract concepts.

.2 CLASSIFICATIONS OF MEASUREMENT SCALES

om what has been stated above, we can write that scales of measurement can be considered in ms of their mathematical properties. The most widely used classification of measurement scales : (a) nominal scale; (b) ordinal scale; (c) interval scale; and (d) ratio scale.

---

[*] When data can be measured in units which are interchangeable e.g., weights (by ratio scales), temperatures (by interval les), that data is said to be parametric and can be subjected to most kinds of statistical and mathematical processes. But en data is measured in units which are not interchangeable, e.g., product preferences (by ordinal scales), the data is said be non-parametric and is susceptible only to a limited extent to mathematical and statistical treatment.

(a) **Nominal scale:** Nominal scale is simply a system of assigning number symbols to eve in order to label them. The usual example of this is the assignment of numbers of basket players in order to identify them. Such numbers cannot be considered to be associated w an ordered scale for their order is of no consequence; the numbers are just conven labels for the particular class of events and as such have no quantitative value. Nom scales provide convenient ways of keeping track of people, objects and events. One can do much with the numbers involved. For example, one cannot usefully average the numb on the back of a group of football players and come up with a meaningful value. Neit can one usefully compare the numbers assigned to one group with the numbers assigne another. The counting of members in each group is the only possible arithmetic operat when a nominal scale is employed. Accordingly, we are restricted to use mode as measure of central tendency. There is no generally used measure of dispersion for nomi scales. Chi-square test is the most common test of statistical significance that can utilized, and for the measures of correlation, the contingency coefficient can be work out.

Nominal scale is the least powerful level of measurement. It indicates no order or distan relationship and has no arithmetic origin. A nominal scale simply describes differen between things by assigning them to categories. Nominal data is, thus, counted data. T scale wastes any information that we may have about varying degrees of attitude, ski understandings, etc. In spite of all this, nominal scales are still very useful and are wid used in surveys and other *ex-post-facto* research when data is being classified by ma sub-groups of the population.

(b) **Ordinal scale:** The lowest level of the ordered scale that is commonly used is the ordi scale. The ordinal scale places events in order, but there is no attempt to make the interv of the scale equal in terms of some rule. Rank orders represent ordinal scales and frequently used in research relating to qualitative phenomena. A student's rank in graduation class involves the use of an ordinal scale. One has to be very careful in mak statement about scores based on ordinal scales. For instance, if Ram's position in his cl is 10 and Mohan's position is 40, it cannot be said that Ram's position is four times as go as that of Mohan. The statement would make no sense at all. Ordinal scales only permit ranking of items from highest to lowest. Ordinal measures have no absolute values, and real differences between adjacent ranks may not be equal. All that can be said is that o person is higher or lower on the scale than another, but more precise comparisons cann be made.

Thus, the use of an ordinal scale implies a statement of 'greater than' or 'less than' (equality statement is also acceptable) without our being able to state how much greater less. The real difference between ranks 1 and 2 may be more or less than the differen between ranks 5 and 6. Since the numbers of this scale have only a rank meaning, t appropriate measure of central tendency is the median. A percentile or quartile measure used for measuring dispersion. Correlations are restricted to various rank order metho Measures of statistical significance are restricted to the non-parametric methods.

(c) **Interval scale:** In the case of interval scale, the intervals are adjusted in terms of so rule that have been established as a basis for making the units equal. The units are equ only in so far as one accepts the assumptions on which the rule is based. Interval sca can have an arbitrary zero, but it is not possible to determine for them what may be called an absolute zero or the unique origin. The primary limitation of the interval scale is the lack of a true zero; it does not have the capacity to measure the complete absence of a trait or characteristic. The Fahrenheit scale is an example of an interval scale and shows similarities in what one can and cannot do with it. One can say that an increase in temperature from 30° to 40° involves the same increase in temperature as an increase from 60° to 70°, but one cannot say that the temperature of 60° is twice as warm as the temperature of 30° because both numbers are dependent on the fact that the zero on the scale is set arbitrarily at the temperature of the freezing point of water. The ratio of the two temperatures, 30° and 60°, means nothing because zero is an arbitrary point.

Interval scales provide more powerful measurement than ordinal scales for interval scale also incorporates the concept of equality of interval. As such more powerful statistical measures can be used with interval scales. Mean is the appropriate measure of central tendency, while standard deviation is the most widely used measure of dispersion. Product moment correlation techniques are appropriate and the generally used tests for statistical significance are the 't' test and 'F' test.

(d) **Ratio scale:** Ratio scales have an absolute or true zero of measurement. The term 'absolute zero' is not as precise as it was once believed to be. We can conceive of an absolute zero of length and similarly we can conceive of an absolute zero of time. For example, the zero point on a centimeter scale indicates the complete absence of length or height. But an absolute zero of temperature is theoretically unobtainable and it remains a concept existing only in the scientist's mind. The number of minor traffic-rule violations and the number of incorrect letters in a page of type script represent scores on ratio scales. Both these scales have absolute zeros and as such all minor traffic violations and all typing errors can be assumed to be equal in significance. With ratio scales involved one can make statements like "Jyoti's" typing performance was twice as good as that of "Reetu." The ratio involved does have significance and facilitates a kind of comparison which is not possible in case of an interval scale.

Ratio scale represents the actual amount of variables. Measures of physical dimensions such as weight, height, distance, etc. are examples. Generally, all statistical techniques are usable with ratio scales and all manipulations that one can carry out with real numbers can also be carried out with ratio scale values. Multiplication and division can be used with this scale but not with other scales mentioned above. Geometric and harmonic means can be used as measures of central tendency and coefficients of variation may also be calculated.

Thus, proceeding from the nominal scale (the least precise type of scale) to ratio scale (the most precise), relevant information is obtained increasingly. If the nature of the variables permits, the researcher should use the scale that provides the most precise description. Researchers in physical sciences have the advantage to describe variables in ratio scale form but the behavioural sciences are generally limited to describe variables in interval scale form, a less precise type of measurement.

5.3 GOODNESS OF MEASUREMENT SCALES

A measurement scale has to have certain desirable qualities to judge their goodness in measuring characteristics under study. These qualities are described below:

(i) *Validity*

Validity is the most critical criterion and indicates the degree to which an instrument measures wl it is supposed to measure. Validity can also be thought of as utility. In other words, validity is extent to which differences found with a measuring instrument reflect true differences among th being tested. But the question arises: how can one determine validity without direct confirmi knowledge? The answer may be that we seek other relevant evidence that confirms the answers have found with our measuring tool. What is relevant, evidence often depends upon the nature of t research problem and the judgement of the researcher. But one can certainly consider three types validity in this connection: (i) Content validity; (ii) Criterion-related validity and (iii) Construct validi

(i) *Content validity* is the extent to which a measuring instrument provides adequate covera of the topic under study. If the instrument contains a representative sample of the univer the content validity is good. Its determination is primarily judgemental and intuitive. It c also be determined by using a panel of persons who shall judge how well the measuri instrument meets the standards, but there is no numerical way to express it.

(ii) *Criterion-related validity* relates to our ability to predict some outcome or estimate t existence of some current condition. This form of validity reflects the success of measur used for some empirical estimating purpose. The concerned criterion must possess t following qualities:

- *Relevance:* (A criterion is relevant if it is defined in terms we judge to be the prop measure.)
- *Freedom from bias:* (Freedom from bias is attained when the criterion gives each subje an equal opportunity to score well.)
- *Reliability:* (A reliable criterion is stable or reproducible.)
- *Availability:* (The information specified by the criterion must be available.)

In fact, a Criterion-related validity is a broad term that actually refers to (i) *Predictive validi* and (ii) *Concurrent validity.* The former refers to the usefulness of a test in predicting some futu performance whereas the latter refers to the usefulness of a test in closely relating to other measur of known validity. Criterion-related validity is expressed as the coefficient of correlation betwee test scores and some measure of future performance or between test scores and scores on anoth measure of known validity.

(iii) *Construct validity* is the most complex and abstract. A measure is said to possess constru validity to the degree that it confirms to predicted correlations with other theoretic propositions. Construct validity is the degree to which scores on a test can be accounte for by the explanatory constructs of a sound theory. For determining construct validity, w associate a set of other propositions with the results received from using our measureme instrument. If measurements on our devised scale correlate in a predicted way with the other propositions, we can conclude that there is some construct validity.

If the above stated criteria and tests are met with, we may state that our measuring instrument alid and will result in correct measurement; otherwise we shall have to look for more information l/or resort to exercise of judgement.

Reliability

e test of reliability is another important test of sound measurement. A measuring instrument is iable if it provides consistent results. Reliable measuring instrument does contribute to validity, but eliable instrument need not be a valid instrument. For instance, a scale that consistently overweighs jects by five kgs., is a reliable scale, but it does not give a valid measure of weight. But the other y is not true i.e., a valid instrument is always reliable. Accordingly, reliability is not as valuable as lidity, but it is easier to assess reliability in comparison to validity. If the quality of reliability is isfied by an instrument, then while using it we can be confident that the transient and situational ctors are not interfering.

Two aspects of reliability viz., stability and equivalence deserve special mention. The *stability pect* is concerned with securing consistent results with repeated measurements of the same person d with the same instrument. We usually determine the degree of stability by comparing the results repeated measurements. The *equivalence aspect* considers how much error may get introduced different investigators or different samples of the items being studied. A good way to test for the uivalence of measurements by two investigators is to compare their observations of the same ents. Reliability can be improved in the following two ways:

(i) By standardising the conditions under which the measurement takes place i.e., we must ensure that external sources of variation such as boredom, fatigue, etc., are minimised to the extent possible. That will improve stability aspect.

(ii) By carefully designed directions for measurement with no variation from group to group, by using trained and motivated persons to conduct the research and also by broadening the sample of items used. This will improve equivalence aspect.

i) Practicality

ne practicality characteristic of a measuring instrument can be judged in terms of economy, nvenience and interpretability. From the operational point of view, the measuring instrument ought be practical i.e., it should be economical, convenient and interpretable. *Economy* consideration ggests that some trade-off is needed between the ideal research project and that which the budget n afford. The length of measuring instrument is an important area where economic pressures are ickly felt. Although more items give greater reliability as stated earlier, but in the interest of limiting e interview or observation time, we have to take only few items for our study purpose. Similarly, ta-collection methods to be used are also dependent at times upon economic factors. *Convenience* st suggests that the measuring instrument should be easy to administer. For this purpose one should ve due attention to the proper layout of the measuring instrument. For instance, a questionnaire, th clear instructions (illustrated by examples), is certainly more effective and easier to complete an one which lacks these features. *Interpretability* consideration is specially important when rsons other than the designers of the test are to interpret the results. The measuring instrument, in der to be interpretable, must be supplemented by (a) detailed instructions for administering the test;) scoring keys; (c) evidence about the reliability and (d) guides for using the test and for interpreting sults.

(iv) *Accuracy*

The characteristic of accuracy of a measurement scale means it should be a true representative the observation of underlying characteristic. For example, measuring with an 'inch' scale will prov accurate value only upto one-eighth of an inch, while measuring with 'cm' scale will provide m accurate value.

5.4 SOURCES OF ERROR IN MEASUREMENT

Measurement should be precise and unambiguous in an ideal research study. This objective, howe is often not met with in entirety. As such the researcher must be aware about the sources of erro measurement. The following are the possible sources of error in measurement.

(a) **Respondent:** At times the respondent may be reluctant to express strong negative feeli or it is just possible that he may have very little knowledge but may not admit his ignoran All this reluctance is likely to result in an interview of 'guesses.' Transient factors li fatigue, boredom, anxiety, etc. may limit the ability of the respondent to respond accurat and fully.

(b) **Situation:** Situational factors may also come in the way of correct measurement. A condition which places a strain on interview can have serious effects on the interview respondent rapport. For instance, if someone else is present, he can distort responses joining in or merely by being present. If the respondent feels that anonymity is not assur he may be reluctant to express certain feelings.

(c) **Measurer:** The interviewer can distort responses by rewording or reordering questio His behaviour, style and looks may encourage or discourage certain replies from responden Careless mechanical processing may distort the findings. Errors may also creep in becau of incorrect coding, faulty tabulation and/or statistical calculations, particularly in the da analysis stage.

(d) **Instrument:** Error may arise because of the defective measuring instrument. The use complex words, beyond the comprehension of the respondent, ambiguous meanings, po printing, inadequate space for replies, response choice omissions, etc. are a few things t make the measuring instrument defective and may result in measurement errors. Anoth type of instrument deficiency is the poor sampling of the universe of items of concern.

Researcher must know that correct measurement depends on successfully meeting all of t problems listed above. He must, to the extent possible, try to eliminate, neutralize or otherwise d with all the possible sources of error so that the final results may not be contaminated.

5.5 TECHNIQUES OF DEVELOPING MEASUREMENT TOOLS

The technique of developing measurement tools involves a four-stage process, consisting of t following:

(a) Concept development;

(b) Specification of concept dimensions;

(c) Selection of indicators; and

(d) Formation of index.

The first and foremost step is that of *concept development* which means that the researcher ould arrive at an understanding of the major concepts pertaining to his study. This step of concept velopment is more apparent in theoretical studies than in the more pragmatic research, where the ndamental concepts are often already established.

The second step requires the researcher to specify the *dimensions of the concepts* that he veloped in the first stage. This task may either be accomplished by deduction i.e., by adopting a ore or less intuitive approach or by empirical correlation of the individual dimensions with the total ncept and/or the other concepts. For instance, one may think of several dimensions such as product putation, customer treatment, corporate leadership, concern for individuals, sense of social sponsibility and so forth when one is thinking about the image of a certain company.

Once the dimensions of a concept have been specified, the researcher must *develop indicators* r measuring each concept element. Indicators are specific questions, scales, or other devices by hich respondent's knowledge, opinion, expectation, etc., are measured. As there is seldom a perfect easure of a concept, the researcher should consider several alternatives for the purpose. The use more than one indicator gives stability to the scores and it also improves their validity.

The last step is that of combining the various indicators into an index, i.e., *formation of an dex*. When we have several dimensions of a concept or different measurements of a dimension, e may need to combine them into a single index. One simple way for getting an overall index is to ovide scale values to the responses and then sum up the corresponding scores. Such an overall dex would provide a better measurement tool than a single indicator. This way we must obtain an verall index for the various concepts concerning the research study.

5.6 SCALING

research we quite often face measurement problem (since we want a valid measurement but may ot obtain it), specially when the concepts to be measured are complex and abstract and we do not ossess the standardised measurement tools. Alternatively, we can say that while measuring attitudes nd opinions, we face the problem of their valid measurement. Similar problem may be faced by a esearcher, of course in a lesser degree, while measuring physical or institutional concepts. As such e should study some procedures which may enable us to measure abstract concepts more accurately. his brings us to the study of scaling techniques.

Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude nd other concepts. This can be done in two ways viz., (i) making a judgement about some characteristic f an individual and then placing him directly on a scale that has been defined in terms of that haracteristic and (ii) constructing questionnaires in such a way that the score of individual's responses ssigns him a place on a scale. It may be stated here that a scale is a continuum, consisting of the ighest point (in terms of some characteristic e.g., preference, favourableness, etc.) and the lowest oint along with several intermediate points between these two extreme points. These scale-point ositions are so related to each other that when the first point happens to be the highest point, the econd point indicates a higher degree in terms of a given characteristic as compared to the third oint and the third point indicates a higher degree as compared to the fourth and so on. Numbers for easuring the distinctions of degree in the attitudes/opinions are, thus, assigned to individuals corresponding to their scale-positions. All this is better understood when we talk about scali technique(s). Hence the term 'scaling' is applied to the procedures for attempting to determi quantitative measures of subjective abstract concepts.

5.7 SCALE CLASSIFICATION BASES

The number assigning procedures or the scaling procedures may be broadly classified on one more of the following bases: (a) subject orientation; (b) response form; (c) degree of subjectivi (d) scale properties; (e) number of dimensions and (f) scale construction techniques. We take each of these separately.

(a) **Subject orientation:** Under it a scale may be designed to measure characteristics of t respondent who completes it or to judge the stimulus object which is presented to t respondent. In respect of the former, we presume that the stimuli presented are sufficient homogeneous so that the between-stimuli variation is small as compared to the variati among respondents. In the latter approach, we ask the respondent to judge some specif object in terms of one or more dimensions and we presume that the between-responde variation will be small as compared to the variation among the different stimuli presented respondents for judging.

(b) **Response form:** Under this we may classify the scales as categorical and comparativ Categorical scales are also known as rating scales. These scales are used when a responde scores some object without direct reference to other objects. Under comparative scale which are also known as ranking scales, the respondent is asked to compare two or mo objects. In this sense the respondent may state that one object is superior to the other that three models of pen rank in order 1, 2 and 3. The essence of ranking is, in fact, relative comparison of a certain property of two or more objects.

(c) **Degree of subjectivity:** With this basis the scale data may be based on whether w measure subjective personal preferences or simply make non-preference judgements. the former case, the respondent is asked to choose which person he favours or whic solution he would like to see employed, whereas in the latter case he is simply asked t judge which person is more effective in some aspect or which solution will take fewe resources without reflecting any personal preference.

(d) **Scale properties:** Considering scale properties, one may classify the scales as nomina ordinal, interval and ratio scales. Nominal scales merely classify without indicating orde distance or unique origin. Ordinal scales indicate magnitude relationships of 'more than' 'less than', but indicate no distance or unique origin. Interval scales have both order an distance values, but no unique origin. Ratio scales possess all these features.

(e) **Number of dimensions:** In respect of this basis, scales can be classified as 'unidimensional' and 'multidimensional' scales. Under the former we measure only one attribute of the respondent or object, whereas multidimensional scaling recognizes that an object might be described better by using the concept of an attribute space of 'n' dimensions, rather than a single-dimension continuum.

(f) **Scale construction techniques:** Following are the five main techniques by which scales can be developed.

(i) *Arbitrary approach:* It is an approach where scale is developed on *adhoc* basis. This is the most widely used approach. It is presumed that such scales measure the concepts for which they have been designed, although there is little evidence to support such an assumption.

(ii) *Consensus approach:* Here a panel of judges evaluate the items chosen for inclusion in the instrument in terms of whether they are relevant to the topic area and unambiguous in implication.

(iii) *Item analysis approach:* Under it a number of individual items are developed into a test which is given to a group of respondents. After administering the test, the total scores are calculated for every one. Individual items are then analysed to determine which items discriminate between persons or objects with high total scores and those with low scores.

(iv) *Cumulative scales* are chosen on the basis of their conforming to some ranking of items with ascending and descending discriminating power. For instance, in such a scale the endorsement of an item representing an extreme position should also result in the endorsement of all items indicating a less extreme position.

(v) *Factor scales* may be constructed on the basis of intercorrelations of items which indicate that a common factor accounts for the relationship between items. This relationship is typically measured through factor analysis method.

5.8 SCALING TECHNIQUES

caling techniques are broadly classified into 'comparative scaling techniques' and 'non-comparative aling techniques'. Comparative scales involve direct comparison of stimulus objects. Comparative ale data are interpreted in relative terms and are measured on ordinal scale. This technique is a on-numeric scaling technique as ordinal data cannot be used for numeric operations. These techniques e easy to understand and apply. They force the respondent to choose between the stimulus objects. or example, respondents are asked whether the like to use toothpaste of brand 'A' or brand 'B'. he respondents have to choose one out of 'A' or 'B' even if there is very small difference in their king of the two brands. The main disadvantage of comparative scales is the inability to generalize yond the stimulus objects. For example, if we want to compare a third brand of toothpaste 'C' with e previous ones, we have to conduct a new study.

In non-comparative scales, each object is scaled independently of the others. For example, the spondents are asked to give preference score on a 1 to 6 scale to brand 'A' of the toothpaste. Here = not preferred at all and 6 = highly preferred. Similar scores can be obtained for brand 'B' and '. Because of the numeric data and wide applications, non-comparative scales are widely used in search.

Now we present some comparative and non-comparative scales which are commonly used.

5.8.1 Comparative Scaling Techniques

Common comparative scaling techniques are

- (i) Paired comparison
- (ii) Rank order
- (iii) Constant sum

A brief discussion on each of the above techniques is given below:

(i) *Paired Comparison*

In this method the respondent can express his attitude by making a choice between two objects, s between a new flavour of soft drink and an established brand of drink. But when there are more th two stimuli to judge, the number of judgements required in a paired comparison is given by t formula:

$$N = \frac{n(n-1)}{2}$$

where N = number of judgements

n = number of stimuli or objects to be judged.

For instance, if there are ten suggestions for bargaining proposals available to a worker's uni there are 45 paired comparisons that can be made with them. When N happens to be a big figu there is the risk of respondents giving ill considered answers or they may even refuse to answer. can reduce the number of comparisons per respondent either by presenting to each one of them o a sample of stimuli or by choosing a few objects which cover the range of attractiveness at ab equal intervals and then comparing all other stimuli to these few standard objects. Thus, pair comparison data may be treated in several ways. If there is substantial consistency, we will find t if X is preferred to Y, and Y to Z, then X will consistently be preferred to Z. If this is true, we may ta the total number of preferences among the comparisons as the score for that stimulus.

It should be remembered that paired comparison provides ordinal data, but the same may converted into an interval scale by the method of the *Law of Comparative Judgement* developed L.L. Thurstone. This technique involves the conversion of frequencies of preferences into a table proportions which are then transformed into Z matrix by referring to the table of area under normal curve. J.P. Guilford in his book "Psychometric Methods" has given a procedure which relatively easier. The method is known as the *Composite Standard Method* and can be illustrated under:

Suppose there are four proposals which some union bargaining committee is considering. The committee wants to know how the union membership ranks these proposals. For this purpose a sample of 100 members might express the views as shown in the following table:

Table 5.1: Response Patterns of 100 Members' Paired Comparisons of 4 Suggestions for Union Bargaining Proposal Priorities

	Suggestion			
	A	*B*	*C*	*D*
A	–	65	32	20
B	40	–	38	42
C	45	50	–	70
D	80	20	98	–
TOTAL:	165	135	168	132
Rank order	2	3	1	4
M_p	0.5375	0.4625	0.5450	0.4550
Z_j	0.09	(–).09	0.11	(–).11
R_j	0.20	0.02	0.22	0.00

Comparing the total number of preferences for each of the four proposals, we find that *C* is the ost popular, followed by *A*, *B* and *D* respectively in popularity. The rank order shown in the above ble explains all this.

By following the composite standard method, we can develop an interval scale from the paired-mparison ordinal data given in the above table for which purpose we have to adopt the following eps in order:

(i) Using the data in the above table, we work out the column mean with the help of the formula given below:

$$M_p = \frac{C + .5(N)}{nN} = \frac{165 + .5(100)}{4(100)} = .5375$$

where

M_p = the mean proportion of the columns

C = the total number of choices for a given suggestion

n = number of stimuli (proposals in the given problem)

N = number of items in the sample.

The column means have been shown in the M_p row in the above table.

(ii) The *Z* values for the M_p are secured from the table giving the area under the normal curve. When the M_p value is less than .5, the *Z* value is negative and for all M_p values higher than .5, the *Z* values are positive.*These *Z* values are shown in Z_j row in the above table.

---

*To use normal curve area table for this sort of transformation, we must subtract 0.5 from all M_p values which exceed to secure the values with which to enter the normal curve area table for which *Z* values can be obtained. For all M_p values less than .5 we must subtract all such values from 0.5 to secure the values with which to enter the normal curve area table which *Z* values can be obtained but the *Z* values in this situation will be with negative sign.

(iii) As the Z_j values represent an interval scale, zero is an arbitrary value. Hence, we c eliminate negative scale values by giving the value of zero to the lowest scale value (th being (–).11 in our example which we shall take equal to zero) and then adding the absolu value of this lowest scale value to all other scale items. This scale has been shown in row in the above table.

Graphically, we can show this interval scale that we have derived from the paired-comparis data using the composite standard method as follows:

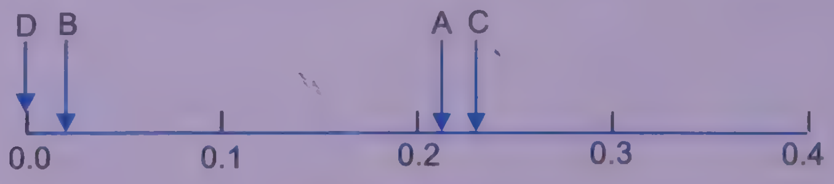


Fig. 5.1

(ii) *Rank Order*

Under this method of comparative scaling, the respondents are asked to rank their choices. Th method is easier and faster than the method of paired comparisons stated above. For example, wi 10 items it takes 45 pair comparisons to complete the task, whereas the method of rank order simp requires ranking of 10 items only. The problem of transitivity (such as *A* prefers to *B*, *B* to *C*, but prefers to *A*) is also not there in case we adopt method of rank order. Moreover, a complete ranki at times is not needed in which case the respondents may be asked to rank only their first, say, fo choices while the number of overall items involved may be more than four, say, it may be 15 or 20 more. To secure a simple ranking of all items involved we simply total rank values received by ea item. There are methods through which we can as well develop an interval scale of these data. B then there are limitations of this method. The first one is that data obtained through this method ordinal data and hence rank ordering is an ordinal scale with all its limitations. Then there may be t problem of respondents becoming careless in assigning ranks particularly when there are ma (usually more than 10) items.

(iii) *Constant Sum*

Constant sum scaling technique is used to assess the relative importance attached by a respondent the stimulus objects. The respondent gives certain points to each stimulus object out of a fixed sum points. This fixed sum is usually taken as 100, but it could be some other value also. For example family planing a holiday, fixes the budget of Rs. 50,000. They wish to plan the expenditures transportation, accommodation, food, drinks, and others. The constant sum is 50,000 in this examp which could be divided as

Transportation	10,000
Accommodation	20,000
Food	12,000
Drink	0
Others	8,000

The data obtained here is numeric but cannot be generalized beyond the list of stimulus objects. ., this data is considered as ordinal data. This technique distinguishes the objects in less time. owever, it is not useful with uneducated people or with large number of objects.

8.2 Non-comparative Scaling Techniques

ost common non-comparative scaling techniques are

(i) Continuous rating or graphic rating
(ii) Itemized rating
(iii) Simple/multiple category scale
(iv) Verbal frequency scale

These techniques are discussed in brief as follows:

) Continuous Rating or Graphic Rating

he graphic rating scale is quite simple and is commonly used in practice. Under it the various oints are usually put along the line to form a continuum and the rater indicates his rating by simply aking a mark (such as √) at the appropriate point on a line that runs from one extreme to the other. cale-points with brief descriptions may be indicated along the line, their function being to assist the ter in performing his job. The following is an example of five-points graphic rating scale when we ish to ascertain people's liking or disliking any product:

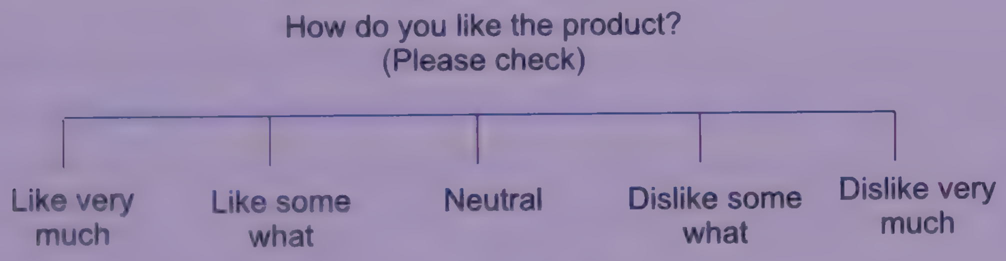


Fig. 5.2

This type of scale has several limitations. The respondents may check at almost any position ong the line which is fact may increase the difficulty of analysis. The meanings of the terms like ery much" and "some what" may depend upon respondent's frame of reference so much so that e statement might be challenged in terms of its equivalency. Several other rating scale variants .g., boxes replacing line) may also be used.

i) Itemized Rating

scale having numbers or brief descriptions of each category is provided. Categories are ordered in rms of scale positions. The respondents select one of the categories that best describes the stimulus oject. Commonly used itemized rating scales are

) Likert Scales

ikert scales or summated scales are developed by utilizing the item analysis approach wherein a articular item is evaluated on the basis of how well it discriminates between those persons whose tal score is high and those whose score is low. Those items or statements that best meet this sort of scrimination test are included in the final instrument.

Thus, summated scales consist of a number of statements which express either a favourable unfavourable attitude towards the given object to which the respondent is asked to react. The respond indicates his agreement or disagreement with each statement in the instrument. Each response given a numerical score, indicating its favourableness or unfavourableness, and the scores are totall to measure the respondent's attitude. In other words, the overall score represents the responden position on the continuum of favourable-unfavourableness towards an issue.

Most frequently used summated scales in the study of social attitudes follow the pattern devis by Likert. For this reason they are often referred to as Likert-type scales. In a Likert scale, t respondent is asked to respond to each of the statements in terms of several degrees, usually fi degrees (but at times 3 or 7 may also be used) of agreement or disagreement. For example, wh asked to express opinion whether one considers his job quite pleasant, the respondent may respond any one of the following ways: (i) strongly agree, (ii) agree, (iii) undecided, (iv) disagree, (v) strong disagree.

We find that these five points constitute the scale. At one extreme of the scale there is stro agreement with the given statement and at the other, strong disagreement, and between them intermediate points. We may illustrate this as under:

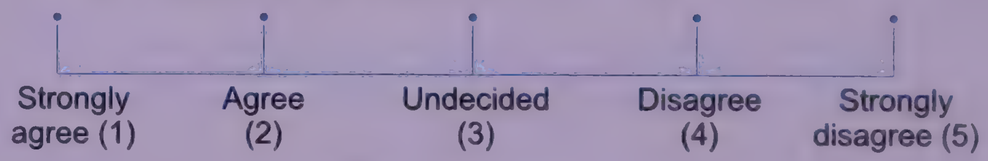


Fig. 5.3

Each point on the scale carries a score. Response indicating the least favourable degree of j satisfaction is given the least score (say 1) and the most favourable is given the highest score (say These score—values are normally not printed on the instrument but are shown here just to indica the scoring pattern. The Likert scaling technique, thus, assigns a scale value to each of the fi responses. The same thing is done in respect of each and every statement in the instrument. Th way the instrument yields a total score for each respondent, which would then measure the respondent favourableness toward the given point of view. If the instrument consists of, say 30 statements, t following score values would be revealing.

$30 \times 5 = 150$ Most favourable response possible

$30 \times 3 = 90$ A neutral attitude

$30 \times 1 = 30$ Most unfavourable attitude.

The scores for any individual would fall between 30 and 150. If the score happens to be abo 90, it shows favourable opinion to the given point of view, a score of below 90 would mean unfavourab opinion and a score of exactly 90 would be suggestive of a neutral attitude.

Procedure: The procedure for developing a Likert-type scale is as follows:

(i) As a first step, the researcher collects a large number of statements which are relevant the attitude being studied and each of the statements expresses definite favourableness unfavourableness to a particular point of view or the attitude and that the number favourable and unfavourable statements is approximately equal.

(ii) After the statements have been gathered, a trial test should be administered to a number of subjects. In other words, a small group of people, from those who are going to be studied finally, are asked to indicate their response to each statement by checking one of the categories of agreement or disagreement using a five point scale as stated above.

(iii) The response to various statements are scored in such a way that a response indicative of the most favourable attitude is given the highest score of 5 and that with the most unfavourable attitude is given the lowest score, say, of 1.

(iv) Then the total score of each respondent is obtained by adding his scores that he received for separate statements.

(v) The next step is to array these total scores and find out those statements which have a high discriminatory power. For this purpose, the researcher may select some part of the highest and the lowest total scores, say the top 25 per cent and the bottom 25 per cent. These two extreme groups are interpreted to represent the most favourable and the least favourable attitudes and are used as criterion groups by which to evaluate individual statements. This way we determine which statements consistently correlate with low favourability and which with high favourability.

(vi) Only those statements that correlate with the total test should be retained in the final instrument and all others must be discarded from it.

dvantages: The Likert-type scale has several advantages. Mention may be made of the important nes.

(a) It is relatively easy to construct the Likert-type scale in comparison to Thurstone-type scale because Likert-type scale can be performed without a panel of judges.

(b) Likert-type scale is considered more reliable because under it respondents answer each statement included in the instrument. As such it also provides more information and data than does the Thurstone-type scale.

(c) Each statement, included in the Likert-type scale, is given an empirical test for discriminating ability and as such, unlike Thurstone-type scale, the Likert-type scale permits the use of statements that are not manifestly related (to have a direct relationship) to the attitude being studied.

(d) Likert-type scale can easily be used in respondent-centred and stimulus-centred studies i.e., through it we can study how responses differ between people and how responses differ between stimuli.

(e) Likert-type scale takes much less time to construct, it is frequently used by the students of opinion research. Moreover, it has been reported in various research studies that there is high degree of correlation between Likert-type scale and Thurstone-type scale.

imitations: There are several limitations of the Likert-type scale as well. One important limitation s that, with this scale, we can simply examine whether respondents are more or less favourable to a opic, but we cannot tell how much more or less they are. There is no basis for belief that the five positions indicated on the scale are equally spaced. The interval between 'strongly agree' and 'agre may not be equal to the interval between "agree" and "undecided". This means that Likert sca does not rise to a stature more than that of an ordinal scale, whereas the designers of Thursto scale claim the Thurstone scale to be an interval scale. One further disadvantage is that often t total score of an individual respondent has little clear meaning since a given total score can secured by a variety of answer patterns. It is unlikely that the respondent can validly react to a sh statement on a printed form in the absence of real-life qualifying situations. Moreover, there "remai a possibility that people may answer according to what they think they should feel rather than ho they do feel."* This particular weakness of the Likert-type scale is met by using a cumulative sca which we shall take up later in this chapter.

In spite of all the limitations, the Likert-type summated scales are regarded as the most useful a situation wherein it is possible to compare the respondent's score with a distribution of scores fro some well defined group. They are equally useful when we are concerned with a programme change or improvement in which case we can use the scales to measure attitudes before and aft the programme of change or improvement in order to assess whether our efforts have had th desired effects. We can as well correlate scores on the scale to other measures without any conce for the absolute value of what is favourable and what is unfavourable. All this accounts for th popularity of Likert-type scales in social studies relating to measuring of attitudes.

(B) *Semantic Differential Scale*

Semantic differential scale or the S.D. scale developed by Charles E. Osgood, G.J. Suci and P.H Tannenbaum (1957), is an attempt to measure the psychological meanings of an object to an individua This scale is based on the presumption that an object can have different dimensions of connotati meanings which can be located in multidimensional property space, or what can be called the semant space in the context of S.D. scale. This scaling consists of a set of bipolar rating scales, usually of points, by which one or more respondents rate one or more concepts on each scale item. For instanc the S.D. scale items for analysing candidates for leadership position may be shown as under:

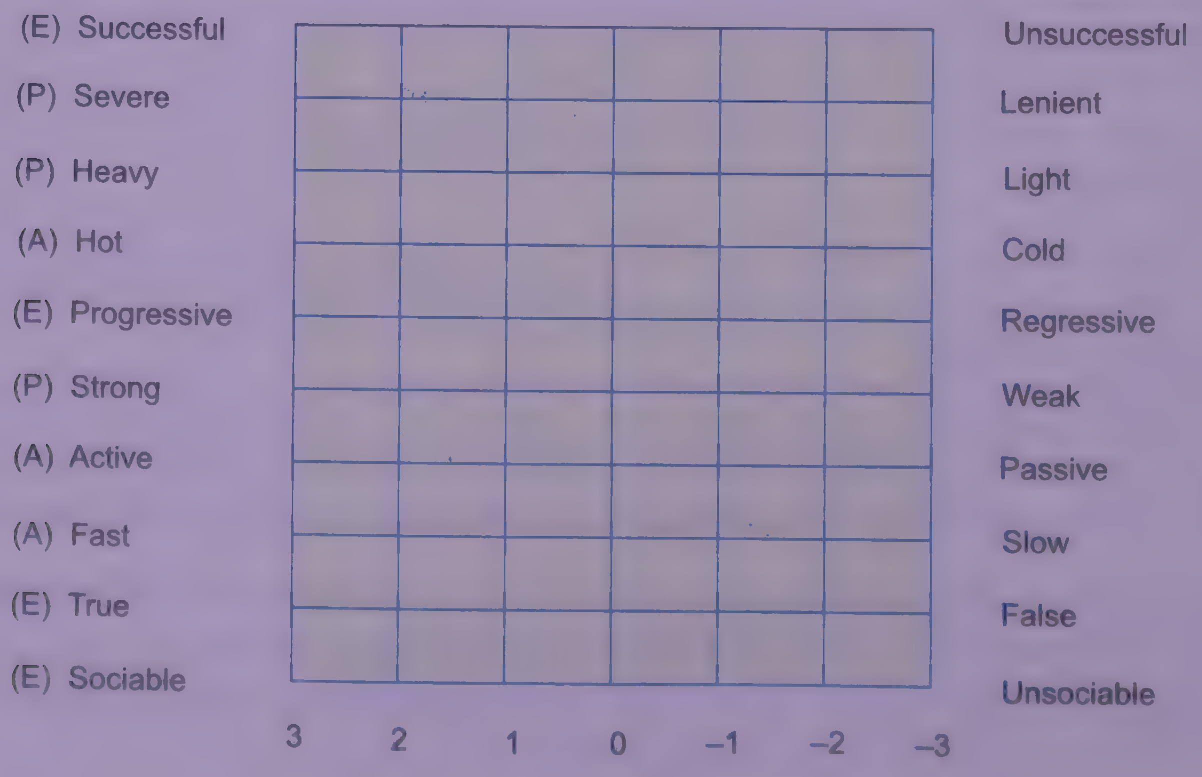


Fig. 5.4

---

* John W. Best and James V. Kahn, *Research in Education*, 5th ed., Prentice-Hall of India Pvt. Ltd., New Delhi, 1986, p. 18

Candidates for leadership position (along with the concept—the 'ideal' candidate) may be ompared and we may score them from +3 to –3 on the basis of the above stated scales. (The tters. *E*, *P*, *A* showing the relevant factor viz., evaluation, potency and activity respectively, written ong the left side are not written in actual scale. Similarly the numeric values shown are also not ritten in actual scale.)

Osgood and others did produce a list of some adjective pairs for attitude research purposes and oncluded that semantic space is multidimensional rather than unidimensional. They made sincere fforts and ultimately found that three factors, viz., evaluation, potency and activity, contributed most o meaningful judgements by respondents. The evaluation dimension generally accounts for 1/2 and /4 of the extractable variance and the other two factors account for the balance.

rocedure: Various steps involved in developing S.D. scale are as follows:

(a) First of all the concepts to be studied are selected. The concepts are usually chosen by personal judgement, keeping in view the nature of the problem.

(b) The next step is to select the scales bearing in mind the criterion of factor composition and the criterion of scale's relevance to the concepts being judged (it is common practice to use at least three scales for each factor with the help of which an average factor score has to be worked out). One more criterion to be kept in view is that scales should be stable across subjects and concepts.

(c) Then a panel of judges are used to rate the various stimuli (or objects) on the various selected scales and the responses of all judges would then be combined to determine the composite scaling.

To conclude, "the S.D. has a number of specific advantages. It is an efficient and easy way to secure attitudes from a large sample. These attitudes may be measured in both direction and intensity. The total set of responses provides a comprehensive picture of the meaning of an object, as well as a measure of the subject doing the rating. It is a standardised technique that is easily repeated, but escapes many of the problems of response distortion found with more direct methods."

(C) *Stapel Scale*

This scale was developed by John Stapel. This is a unipolar rating scale with usually 10 categories numbered from –5 to +5. This scale does not have zero or the neutral point. Respondents rate how each term describes the object by selecting the appropriate number. Positive number means the term describes the object accurately, while negative number implies that the term describes the object inaccurately. +5 means the highest degree of accuracy while –5 means the highest degree of in accuracy.

For example, consider the phrases (*i*) Tasty food, (*ii*) Fast Service, and (*iii*) Good ambience for a restaurant. A respondent is asked to rate how accurately these terms or phrases describe a specified restaurant.

+5	+5	+5
+4	+4	+4
+3	+3	[✓ +3]
+2	+2	+2
[✓ +1]	+1	+1
Tasty Food	Fast Service	Good Ambience
–1	–1	–1
–2	[✓ –2]	–2
–3	–3	–3
–4	–4	–4
–5	–5	–5

The respondent selects +1 for 'tasty food,' – 2 for 'fast service' and +3 for 'good ambience' This means the respondent thinks that the phrase 'tasty food' is accurate with a minimum degree o accuracy, phrase 'fast service' is some what in accurate, while 'good ambience' is sufficiently accurate. So according to the respondent the ambience of the restaurant is good, food tast is okay but service is slow.

This method is applicable when the responses are rated on a single dimension. The method is very economic and data can be collected over telephonic interview also. Some researches think that the method is confusing and of not much use.

(iii) *Single/Multiple Category Scales*

These scales are also known as dichotomous scales. Here we have two or more mutually exclusive responses. For example, – 'Yes' and 'No', 'True' and 'False'. The respondent has to choose only one out of the given categories. It would be clear from the following example questions of a questionnaire:

1. Do you play cricket ?
   ○ Yes ○ No
2. What is your marital status?
   ○ Unmarried ○ Married
   ○ Divorce ○ Widower
3. What is your employment type?
   ○ Salaried ○ Self employed business
   ○ Self employed professional ○ Student
   ○ Retired ○ Home maker

These scales can also be generalized to have more than one responses. For example,

1. What do you own?

    Car ☐
    Computer ☐
    House ☐
    Home theatre system ☐

2. What are you interested in?

    Music ☐
    Sports ☐
    Travelling ☐
    Films ☐
    Reading ☐

3. What do you like to do on internet?

    Social networking ☐
    News reading ☐
    Gaming ☐
    E-mail ☐
    Knowledge gaining ☐

It should be ensured that the provided choice exhaust almost all possible answers of the asked question. The choice 'others' may be given along with other choices. The collected data is on nominal scale. The method is easy and very popular on internet surveys.

(iv) Verbal Frequency Scale

This scale is used when the respondent is unable or unwilling to give the exact numbers in the answer.

For example,
How often do you eatout?

1. Frequently
2. Sometimes
3. Rarely
4. Never

This scale provides only an approximation of frequency and so the data is on ordinal scale.

Dimensions: In unidimensional scale, only one attribute of the object is measured. While, a multidimensional scale considers that an object is described with several dimensions. For example, the popularity of a restaurant can be measured by a single measure food taste. It can be defined on multiple dimensions like food taste, service, cleanliness, ambience, etc.

Number of scale categories: Larger number of scale categories will provide better precision in the collected data. However, it would become inconvinient for the respondent. So, the number of scale categories is a trade-off between precision and convinience. It should be chosen with care.

Balanced and unbalanced scales: In balanced scale, number of favorable and unfavorable categories are equal, while they are unequal in an unbalanced scale. The scale should be balanced in general to collect an unbiased opinion. However, if the distribution of responses is likely to be skewed on one direction, an unbalanced scale having more categories on the opposite direction is used.

5.9 MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) is relatively more complicated scaling device, but with this sort scaling one can scale objects, individuals or both with a minimum of information. Multidimension scaling (or MDS) can be characterized as a set of procedures for portraying perceptual or affectiv dimensions of substantive interest. MDS is used when all the variables (whether metric or no metric) in a study are to be analyzed simultaneously and all such variables happen to be independen The underlying assumption in MDS is that the respondents perceive a set of objects as being more less similar to one another on a number of usually uncorrelated dimensions. Through MDS techniqu one can represent geometrically the locations and interrelationships among a set of points. In fac these techniques attempt to locate the points, given the information about a set of interpoint distance in space of one or more dimensions such as to best summarise the information contained in th interpoint distances. The distances in the solution space then optimally reflect the distances containe in the input data. For instance, if objects, say *X* and *Y*, are thought of by the respondent as being mos similar as compared to all other possible pairs of objects, MDS techniques will position objects *X* an *Y* in such a way that the distance between them in multidimensional space is shorter than tha between any two other objects.

Two approaches, viz., the metric approach and the non-metric approach, are usually talked abou in the context of MDS, while attempting to construct a space containing m points such tha $m(m-1)/2$ interpoint distances reflect the input data. The *metric approach to MDS* treats the inpu data as interval scale data and solves applying statistical methods for the additive constant* which minimises the dimensionality of the solution space. This approach utilises all the information in the data in obtaining a solution. The data (i.e., the metric similarities of the objects) are often obtained on a bipolar similarity scale on which pairs of objects are rated one at a time. If the data reflect exac distances between real objects in an r-dimensional space, their solution will reproduce the set o interpoint distances. But as the true and real data are rarely available, we require random an systematic procedures for obtaining a solution. Generally, the judged similarities among a set of objects are statistically transformed into distances by placing those objects in a multidimensional space of some dimensionality.

The *non-metric approach* first gathers the non-metric similarities by asking respondents to rank order all possible pairs that can be obtained from a set of objects. Such non-metric data is then transformed into some arbitrary metric space and then the solution is obtained by reducing the dimensionality. The non-metric approach has come into prominence during the sixties with the coming into existence of high speed computers to generate metric solutions for ordinal input data.

---

* Additive constant refers to that constant with which one can, either by subtracting or adding, convert interval scale to a ratio scale. For instance, suppose we know that distances, say $a-b$, $b-c$, $c-d$ among stimuli on a ratio scale are 7, 6 and 3 respectively. If one were to subtract 3 from each of these distances, they would be 4, 3 and 0 respectively. The converted distances would be on an interval scale of measurement, but not on a ratio scale. Obviously, one can add 3 to all the converted distances and reachieve the ratio scale of distances. Thus 3 will be taken as the additive constant in this case. Well defined iterative approach is employed in practice for estimating appropriate additive constant.

The MDS techniques, infact, do away with the need in the data collection process to specify the tribute(s) along which the several brands, say of a particular product, may be compared as ultimately e MDS analysis itself reveals such attribute(s) that presumably underlie the expressed relative milarities among objects.

In spite of all the merits stated above, the MDS is not widely used because of the computation omplications involved under it. Many of its methods are quite laborious in terms of both the collection f data and the subsequent analyses. However, some progress has been achieved during the last few ears in the use of non-metric MDS in the context of market research problems.

5.10 DECIDING THE SCALE

efore deciding the scales following factors should be taken into account:

i) Data Type

ach data type can be used for only selected type of analysis. For example, nominal and ordinal data annot be used for many basic statistical analysis. The research should know in advance what kind f analysis will be conducted and what type of data will be required for that analysis. The decision of cale will depend on that data type.

ii) Neutral Category

f it is likely to get neutral or impartial response from some respondents, the neutral category may lso be used.

SUMMARY

We discussed two type of measures 1. qualitative and 2. Quantitative and four types of measurement scales (a) nominal, (b) ordinal, (c) interval and (d) ratio. Measurement scales can be classified into two categories (i) comparative scales and (ii) non-comparative scales. We discussed some scaling techniques under the both categories.

PROBLEMS

1. Describe qualitative and quantitative measures with example.
2. Discuss the four types of measurement scales.
3. Point out the possible sources of measurement error.
4. Explain the criteria of goodness of a measurement scale.
5. Are the following nominal, ordinal, interval or ratio data? Explain your answers.
    - (a) Temperatures measured on the Kelvin scale.
    - (b) Military ranks.
    - (c) Social security numbers.
    - (d) Number of passengers on buses from Delhi to Mumbai.
    - (e) Code numbers given to the religion of persons attempting suicide.

6. The following table shows the results of a paired-comparison preference test of four cold drinks from sample of 200 persons:

Name	*Coca Cola*	*Limca*	*Goldspot*	*Thumps up*
Coca Cola	–	60*	105	45
Limca	160	–	150	70
Goldspot	75	40	–	65
Thumps up	165	120	145	–

* To be read as 60 persons preferred Limca over Coca Cola.

(a) How do these brands rank in overall preference in the given sample.

(b) Develop an interval scale for the four varieties of cold drinks.

7. (a) Narrate the procedure for developing a scalogram and illustrate the same by an example.

(b) Workout Guttman's coefficient of reproducibility from the following information:

Number of cases (N) = 30
Number of items (n) = 6
Number of errors (e) = 10

Interpret the meaning of coefficient you work out in this example.

8. Write short notes on :
   - (a) Likert Scale
   - (b) Semantic Differential Scale
   - (c) Stapel Scale

6

Chapter

Data Collection

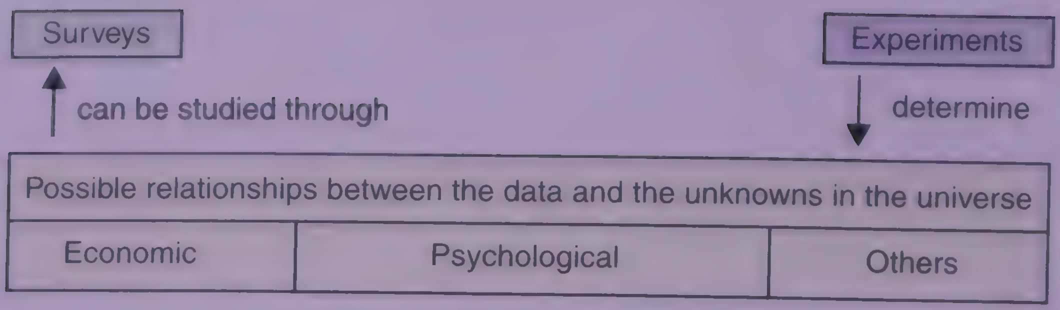
6.1 INTRODUCTION

he task of data collection begins after a research problem has been defined and research design/ lan chalked out. While deciding about the method of data collection to be used for the study, the esearcher should keep in mind two types of data viz., primary and secondary. The *primary data* are hose which are collected afresh and for the first time, and thus happen to be original in character. he *secondary data,* on the other hand, are those which have already been collected by someone lse and which have already been passed through the statistical process. The researcher would have o decide which sort of data he would be using (thus collecting) for his study and accordingly he will ave to select one or the other method of data collection. The methods of collecting primary and econdary data differ since primary data are to be originally collected, while in case of secondary ata the nature of data collection work is merely that of compilation. We describe the different nethods of data collection, with the pros and cons of each method.

6.2 EXPERIMENTS AND SURVEYS

Ve collect primary data during the course of doing experiments in an experimental research. An experiment refers to an investigation in which a factor or variable under test is isolated and its effect(s) measured. In an experiment the investigator measures the effects of an experiment which he conducts intentionally.

However, in case we do research of the descriptive type and perform surveys, whether sample urveys or census surveys, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews. Survey efers to the method of securing information concerning a phenomena under study from all or a selected number of respondents of the concerned universe. In a survey, the investigator examines those phenomena which exist in the universe independent of his action. The difference between an experiment and a survey can be depicted as under:



The following points are noteworthy so far as difference between survey and experiment is concerne

1. Surveys are conducted in case of descriptive research studies where as experiments are part of experimental research studies.
2. Survey-type research studies usually have larger samples because the percentage responses generally happens to be low, as low as 20 to 30%, especially in mailed questionnai studies. Thus, the survey method gathers data from a relatively large number of cases at particular time; it is essentially cross-sectional. As against this, experimental studies generall need small samples.
3. Surveys are concerned with describing, recording, analysing and interpreting condition that either exist or existed. The researcher does not manipulate the variable or arrange f events to happen. Surveys are only concerned with conditions or relationships that exis opinions that are held, processes that are going on, effects that are evident or trends th are developing. They are primarily concerned with the present but at times do conside past events and influences as they relate to current conditions. Thus, in surveys, variable that exist or have already occurred are selected and observed.

   Experimental research provides a systematic and logical method for answering the question "What will happen if this is done when certain variables are carefully controlled o manipulated?" In fact, deliberate manipulation is a part of the experimental method. In a experiment, the researcher measures the effects of an experiment which he conduct intentionally.
4. Surveys are usually appropriate in case of social and behavioural sciences (because man types of behaviour that interest the researcher cannot be arranged in a realistic setting where as experiments are mostly an essential feature of physical and natural sciences.
5. Surveys are an example of field research where as experiments generally constitute a example of laboratory research.
6. Surveys are concerned with hypothesis formulation and testing the analysis of the relationshi between non-manipulated variables. Experimentation provides a method of hypothesis testing After experimenters define a problem, they propose a hypothesis. They then test the hypothesis and confirm or disconfirm it in the light of the controlled variable relationshi that they have observed. The confirmation or rejection is always stated in terms of probability rather than certainty. Experimentation, thus, is the most sophisticated, exacting and powerful method for discovering and developing an organised body of knowledge.
7. Surveys may either be census or sample surveys. They may also be classified as social surveys, economic surveys or public opinion surveys. Whatever be their type, the method of data collection happens to be either observation, or interview or questionnaire/opinionnaire

or some projective technique(s). Case study method can as well be used. But in case of experiments, data are collected from several readings of experiments.

8. In case of surveys, research design must be rigid, must make enough provision for protection against bias and must maximise reliability as the aim happens to be to obtain complete and accurate information. Research design in case of experimental studies, apart reducing bias and ensuring reliability, must permit drawing inferences about causality.
9. Possible relationships between the data and the unknowns in the universe can be studied through surveys where as experiments are meant to determine such relationships.
10. Causal analysis is considered relatively more important in experiments where as in most social and business surveys our interest lies in understanding and controlling relationships between variables and as such correlation analysis is relatively more important in surveys.

.3 COLLECTION OF PRIMARY DATA

ere are several methods of collecting primary data, particularly in surveys and descriptive researches. me important methods are described below:

Observation Method

e observation method is the most commonly used method specially in studies relating to behavioural iences. In a way we all observe things around us, but this sort of observation is not scientific servation. Observation becomes a scientific tool and the method of data collection for the researcher, hen it serves a formulated research purpose, is systematically planned and recorded and is subjected checks and controls on validity and reliability. Under the observation method, the information is ught by way of investigator's own direct observation without asking from the respondent. For stance, in a study relating to consumer behaviour, the investigator instead of asking the brand of rist watch used by the respondent, may himself look at the watch. The main advantage of this ethod is that subjective bias is eliminated, if observation is done accurately. Secondly, the information tained under this method relates to what is currently happening; it is not complicated by either the st behaviour or future intentions or attitudes. Thirdly, this method is independent of respondents' illingness to respond and as such is relatively less demanding of active cooperation on the part of spondents as happens to be the case in the interview or the questionnaire method. This method is rticularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of ving verbal reports of their feelings for one reason or the other.

However, observation method has various limitations. Firstly, it is an expensive method. Secondly, e information provided by this method is very limited. Thirdly, sometimes unforeseen factors may terfere with the observational task. At times, the fact that some people are rarely accessible to rect observation creates obstacle for this method to collect data effectively.

While using this method, the researcher should keep in mind things like: What should be observed? ow the observations should be recorded? Or how the accuracy of observation can be ensured? In se the observation is characterised by a careful definition of the units to be observed, the style of cording the observed information, standardised conditions of observation and the selection of pertinent ta of observation, then the observation is called as *structured observation*. But when observation is to take place without these characteristics to be thought of in advance, the same is termed *unstructured observation.* Structured observation is considered appropriate in descriptive studi whereas in an exploratory study the observational procedure is most likely to be relatively unstructure

We often talk about participant and non-participant types of observation in the context of studi particularly of social sciences. This distinction depends upon the observer's sharing or not shari the life of the group he is observing. If the observer observes by making himself, more or less member of the group he is observing so that he can experience what the members of the gro experience, the observation is called as the *participant observation.* But when the observer observ as a detached emissary without any attempt on his part to experience through participation wh others feel, the observation of this type is often termed as *non-participant observation.* (When t observer is observing in such a manner that his presence may be unknown to the people he observing, such an observation is described as *disguised observation.*)

There are several merits of the participant type of observation: (i) The researcher is enabled record the natural behaviour of the group. (ii) The researcher can even gather information whi could not easily be obtained if he observes in a disinterested fashion. (iii) The researcher can ev verify the truth of statements made by informants in the context of a questionnaire or a schedule. B there are also certain demerits of this type of observation viz., the observer may lose the objectivi to the extent he participates emotionally; the problem of observation-control is not solved; and it ma narrow-down the researcher's range of experience.

Sometimes we talk of *controlled* and *uncontrolled observation.* If the observation takes pla in the natural setting, it may be termed as uncontrolled observation, but when observation takes pla according to definite pre-arranged plans, involving experimental procedure, the same is then term controlled observation. In non-controlled observation, no attempt is made to use precision instrument The major aim of this type of observation is to get a spontaneous picture of life and persons. It has tendency to supply naturalness and completeness of behaviour, allowing sufficient time for observin it. But in controlled observation, we use mechanical (or precision) instruments as aids to accurac and standardisation. Such observation has a tendency to supply formalised data upon whic generalisations can be built with some degree of assurance. The main pitfall of non-controlle observation is that of subjective interpretation. There is also the danger of having the feeling that w know more about the observed phenomena than we actually do. Generally, controlled observatio takes place in various experiments that are carried out in a laboratory or under controlled condition whereas uncontrolled observation is resorted to in case of exploratory researches.

(ii) Interview Method

The interview method of collecting data involves presentation of oral-verbal stimuli and reply i terms of oral-verbal responses. This method can be used through personal interviews and, if possibl through telephone interviews.

(a) *Personal interviews:* Personal interview method requires a person known as the interviewe asking questions generally in a face-to-face contact to the other person or persons. (At times th interviewee may also ask certain questions and the interviewer responds to these, but usually th interviewer initiates the interview and collects the information.) This sort of interview may be in th form of direct personal investigation or it may be indirect oral investigation. In the case of direc rsonal investigation the interviewer has to collect the information personally from the sources ncerned. He has to be on the spot and has to meet people from whom data have to be collected. is method is particularly suitable for intensive investigations. But in certain cases it may not be ssible or worthwhile to contact directly the persons concerned or on account of the extensive ope of enquiry, the direct personal investigation technique may not be used. In such cases an direct oral examination can be conducted under which the interviewer has to cross-examine other rsons who are supposed to have knowledge about the problem under investigation and the formation, obtained is recorded. Most of the commissions and committees appointed by government carry on investigations make use of this method.

The method of collecting information through personal interviews is usually carried out in a uctured way. As such we call the interviews as *structured interviews.* Such interviews involve e use of a set of predetermined questions and of highly standardised techniques of recording. Thus, e interviewer in a structured interview follows a rigid procedure laid down, asking questions in a rm and order prescribed. As against it, the *unstructured interviews* are characterised by a flexibility approach to questioning. Unstructured interviews do not follow a system of pre-determined estions and standardised techniques of recording information. In a non-structured interview, the terviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at nes he may omit certain questions if the situation so requires. He may even change the sequence questions. He has relatively greater freedom while recording the responses to include some aspects d exclude others. But this sort of flexibility results in lack of comparability of one interview with other and the analysis of unstructured responses becomes much more difficult and time-consuming an that of the structured responses obtained in case of structured interviews. Unstructured interviews so demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, wever, happens to be the central technique of collecting information in case of exploratory or rmulative research studies. But in case of descriptive studies, we quite often use the technique of ructured interview because of its being more economical, providing a safe basis for generalisation d requiring relatively lesser skill on the part of the interviewer.

We may as well talk about focussed interview, clinical interview and the non-directive interview. *ocussed interview* is meant to focus attention on the given experience of the respondent and its fects. Under it the interviewer has the freedom to decide the manner and sequence in which the estions would be asked and has also the freedom to explore reasons and motives. The main task of e interviewer in case of a focussed interview is to confine the respondent to a discussion of issues ith which he seeks conversance. Such interviews are used generally in the development of potheses and constitute a major type of unstructured interviews. The *clinical interview* is concerned ith broad underlying feelings or motivations or with the course of individual's life experience. The ethod of eliciting information under it is generally left to the interviewer's discretion. In case of *on-directive interview*, the interviewer's function is simply to encourage the respondent to talk out the given topic with a bare minimum of direct questioning. The interviewer often acts as a atalyst to a comprehensive expression of the respondents' feelings and beliefs and of the frame of ference within which such feelings and beliefs take on personal significance.

Despite the variations in interview-techniques, the major advantages and weaknesses of personal terviews can be enumerated in a general way. The chief merits of the interview method are as llows:

1. More information and that too in greater depth can be obtained.
2. Interviewer by his own skill can overcome the resistance, if any, of the respondents; interview method can be made to yield an almost perfect sample of the general populati
3. There is greater flexibility under this method as the opportunity to restructure question always there, specially in case of unstructured interviews.
4. Observation method can as well be applied to recording verbal answers to various questi
5. Personal information can as well be obtained easily under this method.
6. Samples can be controlled more effectively as there arises no difficulty of the miss returns; non-response generally remains very low.
7. The interviewer can usually control which person(s) will answer the questions. This is possible in mailed questionnaire approach. If so desired, group discussions may also held.
8. The interviewer may catch the informant off-guard and thus may secure the most spontane reactions than would be the case if mailed questionnaire is used.
9. The language of the interview can be adopted to the ability or educational level of person interviewed and as such misinterpretations concerning questions can be avoide
10. The interviewer can collect supplementary information about the respondent's perso characteristics and environment which is often of great value in interpreting results.

But there are also certain weaknesses of the interview method. Among the important weakness mention may be made of the following:

1. It is a very expensive method, specially when large and widely spread geographical sam is taken.
2. There remains the possibility of the bias of interviewer as well as that of the responde there also remains the headache of supervision and control of interviewers.
3. Certain types of respondents such as important officials or executives or people in hi income groups may not be easily approachable under this method and to that extent data may prove inadequate.
4. This method is relatively more-time-consuming, specially when the sample is large and calls upon the respondents are necessary.
5. The presence of the interviewer on the spot may over-stimulate the respondent, sometim even to the extent that he may give imaginary information just to make the intervi interesting.
6. Under the interview method the organisation required for selecting, training and supervisi the field-staff is more complex with formidable problems.
7. Interviewing at times may also introduce systematic errors.
8. Effective interview presupposes proper rapport with respondents that would facilitate fr and frank responses. This is often a very difficult requirement.

Pre-requisites and basic tenets of interviewing: For successful implementation of the intervi method, interviewers should be carefully selected, trained and briefed. They should be honest, since hardworking, impartial and must possess the technical competence and necessary practical experien Occasional field checks should be made to ensure that interviewers are neither cheating, nor deviati n instructions given to them for performing their job efficiently. In addition, some provision should be made in advance so that appropriate action may be taken if some of the selected respondents use to cooperate or are not available when an interviewer calls upon them.

In fact, interviewing is an art governed by certain scientific principles. Every effort should be de to create friendly atmosphere of trust and confidence, so that respondents may feel at ease ile talking to and discussing with the interviewer. The interviewer must ask questions properly and elligently and must record the responses accurately and completely. At the same time, the interviewer st answer legitimate question(s), if any, asked by the respondent and must clear any doubt that the er has. The interviewers approach must be friendly, courteous, conversational and unbiased. The erviewer should not show surprise or disapproval of a respondent's answer but he must keep the ection of interview in his own hand, discouraging irrelevant conversation and must make all possible ort to keep the respondent on the track.

Telephone interviews: This method of collecting information consists in contacting respondents telephone itself. It is not a very widely used method, but plays important part in industrial surveys, ticularly in developed regions. The chief merits of such a system are:

1. It is more flexible in comparison to mailing method.
2. It is faster than other methods i.e., a quick way of obtaining information.
3. It is cheaper than personal interviewing method; here the cost per response is relatively low.
4. Recall is easy; callbacks are simple and economical.
5. There is a higher rate of response than what we have in mailing method; the non-response is generally very low.
6. Replies can be recorded without causing embarrassment to respondents.
7. Interviewer can explain requirements more easily.
8. At times, access can be gained to respondents who otherwise cannot be contacted for one reason or the other.
9. No field staff is required.
10. Representative and wider distribution of sample is possible.

But this system of collecting information is not free from demerits. Some of these may be hlighted.

1. Little time is given to respondents for considered answers; interview period is not likely to exceed five minutes in most cases.
2. Surveys are restricted to respondents who have telephone facilities.
3. Extensive geographical coverage may get restricted by cost considerations.
4. It is not suitable for intensive surveys where comprehensive answers are required to various questions.
5. Possibility of the bias of the interviewer is relatively more.
6. Questions have to be short and to the point; probes are difficult to handle.

Interviewing is an art and one learns it by experience. However, the following points may be kept view by an interviewer for eliciting the desired information:

1. Interviewer must plan in advance and should fully know the problem under considerat He must choose a suitable time and place so that the interviewee may be at ease during interview period. For this purpose some knowledge of the daily routine of the intervie is essential.
2. Interviewer's approach must be friendly and informal. Initially friendly greeting accordance with the cultural pattern of the interviewee should be exchanged and then purpose of the interview should be explained.
3. All possible effort should be made to establish proper rapport with the interviewee; pe are motivated to communicate when the atmosphere is favourable.
4. Interviewer must know that ability to listen with understanding, respect and curiosity is gateway to communication, and hence must act accordingly during the interview. For this, the interviewer must be intelligent and must be a man with self-restraint and s discipline.
5. To the extent possible there should be a free-flowing interview and the questions must well phrased in order to have full cooperation of the interviewee. But the interviewer m control the course of the interview in accordance with the objective of the study.
6. In case of big enquiries, where the task of collecting information is to be accomplished several interviewers, there should be an interview guide to be observed by all so as ensure reasonable uniformity in respect of all salient points in the study.

(iii) Collection of Data Through Questionnaires

This method of data collection is quite popular, particularly in case of big enquiries. It is being adop by private individuals, research workers, private and public organisations and even by governmen In this method a questionnaire is sent (usually by post) to the persons concerned with a request answer the questions and return the questionnaire. A questionnaire consists of a number of questio printed or typed in a definite order on a form or set of forms. The questionnaire is mailed to responde who are expected to read and understand the questions and write down the reply in the space me for the purpose in the questionnaire itself. The respondents have to answer the questions on th own.

The method of collecting data by mailing the questionnaires to respondents is most extensiv employed in various economic and business surveys. The merits claimed on behalf of this method as follows:

1. There is low cost even when the universe is large and is widely spread geographically.
2. It is free from the bias of the interviewer; answers are in respondents' own words.
3. Respondents have adequate time to give well thought out answers.
4. Respondents, who are not easily approachable, can also be reached conveniently.
5. Large samples can be made use of and thus the results can be made more dependable a reliable.

The main demerits of this system can also be listed here:

1. Low rate of return of the duly filled in questionnaires; bias due to no-response is often indeterminate.
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.
4. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched.
5. There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

Before using this method, it is always advisable to conduct 'pilot study' (Pilot Survey) for testing questionnaires. In a big enquiry the significance of pilot survey is felt very much. Pilot survey is act the replica and rehearsal of the main survey. Such a survey, being conducted by experts, brings he light the weaknesses (if any) of the questionnaires and also of the survey techniques. From the erience gained in this way, improvement can be effected.

Main aspects of a questionnaire: Quite often questionnaire is considered as the heart of a vey operation. Hence it should be very carefully constructed. If it is not properly set up, then the vey is bound to fail. This fact requires us to study the main aspects of a questionnaire viz., the eral form, question sequence and question formulation and wording. Researcher should note the lowing with regard to these three main aspects of a questionnaire:

General form: So far as the general form of a questionnaire is concerned, it can either be uctured or unstructured questionnaire. Structured questionnaires are those questionnaires in which re are definite, concrete and pre-determined questions. The questions are presented with exactly same wording and in the same order to all respondents. Resort is taken to this sort of standardisation ensure that all respondents reply to the same set of questions. The form of the question may be her closed (i.e., of the type 'yes' or 'no') or open (i.e., inviting free response) but should be stated advance and not constructed during questioning. Structured questionnaires may also have fixed ernative questions in which responses of the informants are limited to the stated alternatives. Thus ighly structured questionnaire is one in which all questions and answers are specified and comments he respondent's own words are held to the minimum. When these characteristics are not present a questionnaire, it can be termed as unstructured or non-structured questionnaire. More specifically, can say that in an unstructured questionnaire, the interviewer is provided with a general guide on type of information to be obtained, but the exact question formulation is largely his own responsibility d the replies are to be taken down in the respondent's own words to the extent possible; in some uations tape recorders may be used to achieve this goal.

Structured questionnaires are simple to administer and relatively inexpensive to analyse. The vision of alternative replies, at times, helps to understand the meaning of the question clearly. But ch questionnaires have limitations too. For instance, wide range of data and that too in respondent's n words cannot be obtained with structured questionnaires. They are usually considered inappropriate nvestigations where the aim happens to be to probe for attitudes and reasons for certain actions or lings. They are equally not suitable when a problem is being first explored and working hypotheses sought. In such situations, unstructured questionnaires may be used effectively. Then on the basi the results obtained in pretest (testing before final use) operations from the use of unstructu questionnaires, one can construct a structured questionnaire for use in the main study.

2. *Question sequence:* In order to make the questionnaire effective and to ensure quality to replies received, a researcher should pay attention to the question-sequence in preparing questionnaire. A proper sequence of questions reduces considerably the chances of individual questi being misunderstood. The question-sequence must be clear and smoothly-moving, meaning ther that the relation of one question to another should be readily apparent to the respondent, with questi that are easiest to answer being put in the beginning. The`first few questions are particularly impor because they are likely to influence the attitude of the respondent and in seeking his desi cooperation. The opening questions should be such as to arouse human interest. The following t of questions should generally be avoided as opening questions in a questionnaire:

(a) questions that put too great a strain on the memory or intellect of the respondent;
(b) questions of a personal character;
(c) questions related to personal wealth, etc.

Following the opening questions, we should have questions that are really vital to the resea problem and a connecting thread should run through successive questions. Ideally, the questi sequence should conform to the respondent's way of thinking. Knowing what information is desir the researcher can rearrange the order of the questions (this is possible in case of unstructu questionnaire) to fit the discussion in each particular case. But in a structured questionnaire the b that can be done is to determine the question-sequence with the help of a Pilot Survey which is lik to produce good rapport with most respondents. Relatively difficult questions must be relega towards the end so that even if the respondent decides not to answer such questions, considera information would have already been obtained. Thus, question-sequence should usually go from general to the more specific and the researcher must always remember that the answer to a giv question is a function not only of the question itself, but of all previous questions as well. For instan if one question deals with the price usually paid for coffee and the next with reason for preferr that particular brand, the answer to this latter question may be couched largely in terms of pri differences.

3. *Question formulation and wording:* With regard to this aspect of questionnaire, the researc should note that each question must be very clear for any sort of misunderstanding can do irrepara harm to a survey. Question should also be impartial in order not to give a biased picture of the t state of affairs. Questions should be constructed with a view to their forming a logical part of a w thought out tabulation plan. In general, all questions should meet the following standards—(a) shou be easily understood; (b) should be simple i.e., should convey only one thought at a time; (c) shou be concrete and should conform as much as possible to the respondent's way of thinking. (F instance, instead of asking. "How many razor blades do you use annually?" The more realis question would be to ask, "How many razor blades did you use last week?"

Concerning the form of questions, we can talk about two principal forms, viz., multiple cho question and the open-end question. In the former the respondent selects one of the alternati possible answers put to him, whereas in the latter he has to supply the answer in his own words. T question with only two possible answers (usually 'Yes' or 'No') can be taken as a special case of t ltiple choice question, or can be named as a 'closed question.' There are some advantages and advantages of each possible form of question. Multiple choice or closed questions have the antages of easy handling, simple to answer, quick and relatively inexpensive to analyse. They are st amenable to statistical analysis. Sometimes, the provision of alternative replies helps to make ar the meaning of the question. But the main drawback of fixed alternative questions is that of utting answers in people's mouths" i.e., they may force a statement of opinion on an issue about ich the respondent does not infact have any opinion. They are not appropriate when the issue der consideration happens to be a complex one and also when the interest of the researcher is in exploration of a process. In such situations, open-ended questions which are designed to permit ee response from the respondent rather than one limited to certain stated alternatives are considered propriate. Such questions give the respondent considerable latitude in phrasing a reply. Getting the lies in respondent's own words is, thus, the major advantage of open-ended questions. But one ould not forget that, from an analytical point of view, open-ended questions are more difficult to ndle, raising problems of interpretation, comparability and interviewer bias.*

In practice, one rarely comes across a case when one questionnaire relies on one form of estions alone. The various forms complement each other. As such questions of different forms are cluded in one single questionnaire. For instance, multiple-choice questions constitute the basis of a uctured questionnaire, particularly in a mail survey. But even there, various open-ended questions e generally inserted to provide a more complete picture of the respondent's feelings and attitudes.

Researcher must pay proper attention to the wordings of questions since reliable and meaningful turns depend on it to a large extent. Since words are likely to affect responses, they should be operly chosen. Simple words, which are familiar to all respondents should be employed. Words th ambiguous meanings must be avoided. Similarly, danger words, catch-words or words with notional connotations should be avoided. Caution must also be exercised in the use of phrases hich reflect upon the prestige of the respondent. Question wording, in no case, should bias the swer. In fact, question wording and formulation is an art and can only be learnt by practice.

Essentials of a good questionnaire: To be successful, questionnaire should be comparatively ort and simple i.e., the size of the questionnaire should be kept to the minimum. Questions should oceed in logical sequence moving from easy to more difficult questions. Personal and intimate estions should be left to the end. Technical terms and vague expressions capable of different terpretations should be avoided in a questionnaire. Questions may be dichotomous (yes or no swers), multiple choice (alternative answers listed) or open-ended. The latter type of questions are ten difficult to analyse and hence should be avoided in a questionnaire to the extent possible. There ould be some control questions in the questionnaire which indicate the reliability of the respondent. r instance, a question designed to determine the consumption of particular material may be asked st in terms of financial expenditure and later in terms of weight. The control questions, thus, troduce a cross-check to see whether the information collected is correct or not. Questions affecting e sentiments of respondents should be avoided. Adequate space for answers should be provided in e questionnaire to help editing and tabulation. There should always be provision for indications of certainty, e.g., "do not know," "no preference" and so on. Brief directions with regard to filling up e questionnaire should invariably be given in the questionnaire itself. Finally, the physical appearance f the questionnaire affects the cooperation the researcher receives from the recipients and as such an attractive looking questionnaire, particularly in mail surveys, is a plus point for enlisting cooperati The quality of the paper, along with its colour, must be good so that it may attract the attentio recipients.

---

* Interviewer bias refers to the extent to which an answer is altered in meaning by some action or attitude on the part of e interviewer.

Nowadays, the usage of internet is quite popular in collecting the data. Questionnaires are through e-mails to randomly or purposely selected people. Their e-mail ids can be taken from sc other surveys. The trend of paid surveys had also started. There are some websites which pay filling the questionnaires.

(iv) Collection of Data Through Schedules

This method of data collection is very much like the collection of data through questionnaire, w little difference which lies in the fact that schedules (proforma containing a set of questions) being filled in by the enumerators who are specially appointed for the purpose. These enumerat along with schedules, go to respondents, put to them the questions from the proforma in the order questions are listed and record the replies in the space meant for the same in the proforma. In cert situations, schedules may be handed over to respondents and enumerators may help them in record their answers to various questions in the said schedules. Enumerators explain the aims and objects the investigation and also remove the difficulties which any respondent may feel in understanding implications of a particular question or the definition or concept of difficult terms.

This method requires the selection of enumerators for filling up schedules or assisting responde to fill up schedules and as such enumerators should be very carefully selected. The enumerat should be trained to perform their job well and the nature and scope of the investigation should explained to them thoroughly so that they may well understand the implications of different questio put in the schedule. Enumerators should be intelligent and must possess the capacity of cro examination in order to find out the truth. Above all, they should be honest, sincere, hardworking a should have patience and perseverance.

This method of data collection is very useful in extensive enquiries and can lead to fairly relial results. It is, however, very expensive and is usually adopted in investigations conducted by governmen agencies or by some big organisations. Population census all over the world is conducted through tl method.

6.3.1 Difference between Questionnaire and Schedule

Both questionnaire and schedule are popularly used methods of collecting data in research survey There is much resemblance in the nature of these two methods and this fact has made many peop to remark that from a practical point of view, the two methods can be taken to be the same. But fro the technical point of view there is difference between the two. The important points of differen are as under:

1. The questionnaire is generally sent through mail to informants to be answered as specifi in a covering letter, but otherwise without further assistance from the sender. The schedu is generally filled out by the research worker or the enumerator, who can interpret questio when necessary.

2. To collect data through questionnaire is relatively cheap and economical since we have to spend money only in preparing the questionnaire and in mailing the same to respondents. Here no field staff required. To collect data through schedules is relatively more expensive since considerable amount of money has to be spent in appointing enumerators and in importing training to them. Money is also spent in preparing schedules.
3. Non-response is usually high in case of questionnaire as many people do not respond and many return the questionnaire without answering all questions. Bias due to non-response often remains indeterminate. As against this, non-response is generally very low in case of schedules because these are filled by enumerators who are able to get answers to all questions. But there remains the danger of interviewer bias and cheating.
4. In case of questionnaire, it is not always clear as to who replies, but in case of schedule the identity of respondent is known.
5. The questionnaire method is likely to be very slow since many respondents do not return the questionnaire in time despite several reminders, but in case of schedules the information is collected well in time as they are filled in by enumerators.
6. Personal contact is generally not possible in case of the questionnaire method as questionnaires are sent to respondents by post who also in turn return the same by post. But in case of schedules direct personal contact is established with respondents.
7. Questionnaire method can be used only when respondents are literate and cooperative, but in case of schedules the information can be gathered even when the respondents happen to be illiterate.
8. Wider and more representative distribution of sample is possible under the questionnaire method, but in respect of schedules there usually remains the difficulty in sending enumerators over a relatively wider area.
9. Risk of collecting incomplete and wrong information is relatively more under the questionnaire method, particularly when people are unable to understand questions properly. But in case of schedules, the information collected is generally complete and accurate as enumerators can remove the difficulties, if any, faced by respondents in correctly understanding the questions. As a result, the information collected through schedules is relatively more accurate than that obtained through questionnaires.
10. The success of questionnaire method lies more on the quality of the questionnaire itself, but in the case of schedules much depends upon the honesty and competence of enumerators.
11. In order to attract the attention of respondents, the physical appearance of questionnaire must be quite attractive, but this may not be so in case of schedules as they are to be filled in by enumerators and not by respondents.
12. Along with schedules, observation method can also be used but such a thing is not possible while collecting data through questionnaires.

.3.2 Guidelines for Constructing Questionnaire/Schedule

he researcher must pay attention to the following points in constructing an appropriate and effective uestionnaire or a schedule:

1. The researcher must keep in view the problem he is to study for it provides the starti point for developing the Questionnaire/Schedule. He must be clear about the various aspe of his research problem to be dealt with in the course of his research project.
2. Appropriate form of questions depends on the nature of information sought, the sampl respondents and the kind of analysis intended. The researcher must decide whether to u closed or open-ended question. Questions should be simple and must be constructed wit view to their forming a logical part of a well thought out tabulation plan. The units enumeration should also be defined precisely so that they can ensure accurate and fu information.
3. Rough draft of the Questionnaire/Schedule be prepared, giving due thought to the appropria sequence of putting questions. Questionnaires or schedules previously drafted (if availabl may as well be looked into at this stage.
4. Researcher must invariably re-examine, and in case of need may revise the rough draft f a better one. Technical defects must be minutely scrutinised and removed.
5. Pilot study should be undertaken for pre-testing the questionnaire. The questionnaire ma be edited in the light of the results of the pilot study.
6. Questionnaire must contain simple but straight forward directions for the respondents s that they may not feel any difficulty in answering the questions.

6.3.3 Some Other Methods of Data Collection

Let us consider some other methods of data collection, particularly used by big business houses i modern times.

1. **Warranty cards:** Warranty cards are usually postal sized cards which are used by dealer of consumer durables to collect information regarding their products. The information sough is printed in the form of questions on the 'warranty cards' which is placed inside th package along with the product with a request to the consumer to fill in the card and post back to the dealer.
2. **Distributor or store audits:** Distributor or store audits are performed by distributors a well as manufactures through their salesmen at regular intervals. Distributors get the reta stores audited through salesmen and use such information to estimate market size, mark share, seasonal purchasing pattern and so on. The data are obtained in such audits not b questioning but by observation. For instance, in case of a grocery store audit, a sample o stores is visited periodically and data are recorded on inventories on hand either b observation or copying from store records. Store audits are invariably panel operations, fo the derivation of sales estimates and compilation of sales trends by stores are their principa '*raison detre*'. The principal advantage of this method is that it offers the most efficien way of evaluating the effect on sales of variations of different techniques of in-stor promotion.
3. **Pantry audits:** Pantry audit technique is used to estimate consumption of the basket o goods at the consumer level. In this type of audit, the investigator collects an inventory o types, quantities and prices of commodities consumed. Thus in pantry audit data are recorde from the examination of consumer's pantry. The usual objective in a pantry audit is to fin

out what types of consumers buy certain products and certain brands, the assumption being that the contents of the pantry accurately portray consumer's preferences. Quite often, pantry audits are supplemented by direct questioning relating to reasons and circumstances under which particular products were purchased in an attempt to relate these factors to purchasing habits. A pantry audit may or may not be set up as a panel operation, since a single visit is often considered sufficient to yield an accurate picture of consumers' preferences. An important limitation of pantry audit approach is that, at times, it may not be possible to identify consumers' preferences from the audit data alone, particularly when promotion devices produce a marked rise in sales.

4. **Consumer panels:** An extension of the pantry audit approach on a regular basis is known as 'consumer panel', where a set of consumers are arranged to come to an understanding to maintain detailed daily records of their consumption and the same is made available to investigator on demands. In other words, a consumer panel is essentially a sample of consumers who are interviewed repeatedly over a period of time. Mostly consume panels are of two types viz., the transitory consumer panel and the continuing consumer panel. *A transitory consumer panel* is set up to measure the effect of a particular phenomenon. Usually such a panel is conducted on a before-and-after-basis. Initial interviews are conducted before the phenomenon takes place to record the attitude of the consumer. A second set of interviews is carried out after the phenomenon has taken place to find out the consequent changes that might have occurred in the consumer's attitude. It is a favourite tool of advertising and of social research. A *continuing consumer panel* is often set up for an indefinite period with a view to collect data on a particular aspect of consumer behaviour over time, generally at periodic intervals or may be meant to serve as a general purpose panel for researchers on a variety of subjects. Such panels have been used in the area of consumer expenditure, public opinion and radio and TV listenership among others. Most of these panels operate by mail. The representativeness of the panel relative to the population and the effect of panel membership on the information obtained after the two major problems associated with the use of this method of data collection.
5. **Use of mechanical devices:** The use of mechanical devices has been widely made to collect information by way of indirect means. Eye camera, Pupilometric camera, Psychogalvanometer, Motion picture camera and Audiometer are the principal devices so far developed and commonly used by modern big business houses, mostly in the developed world for the purpose of collecting the required information.

Eye cameras are designed to record the focus of eyes of a respondent on a specific portion of a ketch or diagram or written material. Such an information is useful in designing advertising material. upilometric cameras record dilation of the pupil as a result of a visual stimulus. The extent of ilation shows the degree of interest aroused by the stimulus. Psychogalvanometer is used for measuring he extent of body excitement as a result of the visual stimulus. Motion picture cameras can be used o record movement of body of a buyer while deciding to buy a consumer good from a shop or big tore. Influence of packaging or the information given on the label would stimulate a buyer to perform ertain physical movements which can easily be recorded by a hidden motion picture camera in the hop's four walls. Audiometers are used by some TV concerns to find out the type of programmes as vell as stations preferred by people. A device is fitted in the television instrument itself to record hese changes. Such data may be used to find out the market share of competing television stations.

6. **Projective techniques:** Projective techniques (or what are sometimes called as indire interviewing techniques) for the collection of data have been developed by psychologists use projections of respondents for inferring about underlying motives, urges, or intentio which are such that the respondent either resists to reveal them or is unable to figure c himself. In projective techniques the respondent in supplying information tends unconscious to project his own attitudes or feelings on the subject under study. Projective techniqu play an important role in motivational researches or in attitude surveys.

The use of these techniques requires intensive specialised training. In such techniques, th individual's responses to the stimulus-situation are not taken at their face value. The stimuli ma arouse many different kinds of reactions. The nature of the stimuli and the way in which they ar presented under these techniques do not clearly indicate the way in which the response is to interpreted. The stimulus may be a photograph, a picture, an inkblot and so on. Responses to the stimuli are interpreted as indicating the individual's own view, his personality structure, his need tensions, etc. in the context of some pre-established psychological conceptualisation of what th individual's responses to the stimulus mean.

We may now briefly deal with the important projective techniques.

(i) *Word association tests:* These tests are used to extract information regarding such word which have maximum association. In this sort of test the respondent is asked to mention th first word that comes to mind, ostensibly without thinking, as the interviewer reads out eac word from a list. If the interviewer says *cold*, the respondent may say *hot* and the lik ones. The general technique is to use a list of as many as 50 to 100 words. Analysis of th matching words supplied by the respondents indicates whether the given word should b used for the contemplated purpose. The same idea is exploited in marketing research t find out the quality that is mostly associated to a brand of a product. A number of qualitie of a product may be listed and informants may be asked to write brand names possessin one or more of these. This technique is quick and easy to use, but yields reliable result when applied to words that are widely known and which possess essentially one type c meaning. This technique is frequently used in advertising research.

(ii) *Sentence completion tests:* These tests happen to be an extension of the technique c word association tests. Under this, informant may be asked to complete a sentence (suc as: persons who wear Khadi are...) to find association of Khadi clothes with certai personality characteristics. Several sentences of this type might be put to the informant o the same subject. Analysis of replies from the same informant reveals his attitude towar that subject, and the combination of these attitudes of all the sample members is then take to reflect the views of the population. This technique permits the testing not only of word (as in case of word association tests), but of ideas as well and thus, helps in developin hypotheses and in the construction of questionnaires. This technique is also quick and eas to use, but it often leads to analytical problems, particularly when the response happens t be multidimensional.

(iii) *Story completion tests:* Such tests are a step further wherein the researcher may contriv stories instead of sentences and ask the informants to complete them. The respondent i given just enough of story to focus his attention on a given subject and he is asked to suppl a conclusion to the story.

(iv) *Verbal projection tests:* These are the tests wherein the respondent is asked to comment on or to explain what other people do. For example, why do people smoke? Answers may reveal the respondent's own motivations.

(v) *Pictorial techniques:* There are several pictorial techniques. The important ones are as follows:

(a) *Thematic apperception test (T.A.T.):* The TAT consists of a set of pictures (some of the pictures deal with the ordinary day-to-day events while others may be ambiguous pictures of unusual situations) that are shown to respondents who are asked to describe what they think the pictures represent. The replies of respondents constitute the basis for the investigator to draw inferences about their personality structure, attitudes, etc.

(b) *Rosenzweig test:* This test uses a cartoon format wherein we have a series of cartoons with words inserted in 'balloons' above. The respondent is asked to put his own words in an empty balloon space provided for the purpose in the picture. From what the respondents write in this fashion, the study of their attitudes can be made.

(c) *Rorschach test:* This test consists of ten cards having prints of inkblots. The design happens to be symmetrical but meaningless. The respondents are asked to describe what they perceive in such symmetrical inkblots and the responses are interpreted on the basis of some pre-determined psychological framework. This test is frequently used but the problem of validity still remains a major problem of this test.

(d) *Holtzman inkblot test (HIT):* This test from W.H. Holtzman is a modification of the Rorschach Test explained above. This test consists of 45 inkblot cards (and not 10 inkblots as we find in case of Rorschach Test) which are based on colour, movement, shading and other factors involved in inkblot perception. Only one response per card is obtained from the subject (or the respondent) and the responses of a subject are interpreted at three levels of form appropriateness. Form responses are interpreted for knowing the accuracy (F) or inaccuracy (F–) of respondent's percepts; shading and colour for ascertaining his affectional and emotional needs; and movement responses for assessing the dynamic aspects of his life.

Holtzman Inkblot Test or H.I.T. has several special features or advantages. For example, it licits relatively constant number of responses per respondent. Secondly, it facilitates studying the esponses of a respondent to different cards in the light of norms of each card instead of lumping hem together. Thirdly, it elicits much more information from the respondent then is possible with nerely 10 cards in Rorschach test; the 45 cards used in this test provide a variety of stimuli to the espondent and as such the range of responses elicited by the test is comparatively wider.

There are some limitations of this test as well. One difficulty that remains in using this test is that nost of the respondents do not know the determinants of their perceptions, but for the researcher, vho has to interpret the protocols of a subject and understand his personality (or attitude) through hem, knowing the determinant of each of his response is a must. This fact emphasises that the test nust be administered individually and a post-test inquiry must as well be conducted for knowing the ature and sources of responses and this limits the scope of HIT as a group test of personality.

In view of these limitations, some people have made certain changes in applying this test. For nstance, Fisher and Cleveland in their approach for obtaining Barrier score of an individual's personality ave developed a series of multiple choice items for 40 of HIT cards. Each of these cards is presented to the subject along with three acceptable choices [such as 'Knight in armour' (Barrier respons 'X-Ray' (Penetrating response) and 'Flower' (Neutral response)]. Subject taking the test is to che the choice he likes most, make a different mark against the one he likes least and leave the th choice blank. The number of barrier responses checked by him determines his barrier score on t test.

(e) *Tomkins-Horn picture arrangement test:* This test is designed for group administratio It consists of twenty-five plates, each containing three sketches that may be arranged different ways to portray sequence of events. The respondent is asked to arrange them a sequence which he considers as reasonable. The responses are interpreted as providi evidence confirming certain norms, respondent's attitudes, etc.

(vi) *Play techniques:* Under play techniques subjects are asked to improvise or act out situation in which they have been assigned various roles. The researcher may obser such traits as hostility, dominance, sympathy, prejudice or the absence of such traits. The techniques have been used for knowing the attitudes of younger ones through manipulati of dolls. Dolls representing different racial groups are usually given to children who a allowed to play with them freely. The manner in which children organise dolls would indica their attitude towards the class of persons represented by dolls. This is also known as *do play test*, and is used frequently in studies pertaining to sociology. The choice of colou form, words, the sense of orderliness and other reactions may provide opportunities to inf deep-seated feelings.

(vii) *Quizzes, tests and examinations:* This is also a technique of extracting informatio regarding specific ability of candidates indirectly. In this procedure both long and sho questions are framed to test through them the memorising and analytical ability of candidate

(viii) *Sociometry:* Sociometry is a technique for describing the social relationships amon individuals in a group. In an indirect way, sociometry attempts to describe attractions o repulsions between individuals by asking them to indicate whom they would choose o reject in various situations. Thus, sociometry is a new technique of studying the underlyin motives of respondents. Sociograms are charts that depict the sociometric choices. Ther are many versions of the sociogram pattern and the reader is suggested to consult specialise references on sociometry for the purpose. This approach has been applied to the diffusio of ideas on drugs amongst medical practitioners.

7. **Depth interviews:** Depth interviews are those interviews that are designed to discove underlying motives and desires and are often used in motivational research. Such interview are held to explore needs, desires and feelings of respondents. In other words, they aim t elicit unconscious as also other types of material relating especially to personality dynamic and motivations. As such, depth interviews require great skill on the part of the intervie and at the same time involve considerable time. Unless the researcher has specialise training, depth interviewing should not be attempted.

Depth interview may be projective in nature or it may be a non-projective interview. The differenc lies in the nature of the questions asked. Indirect questions on seemingly irrelevant subjects provid information that can be related to the informant's behaviour or attitude towards the subject unde study. Thus, for instance, the informant may be asked on his frequency of air travel and he migh ain be asked at a later stage to narrate his opinion concerning the feelings of relatives of some her man who gets killed in an airplane accident. Reluctance to fly can then be related to replies to uestions of the latter nature. If the depth interview involves questions of such type, the same may be eated as projective depth interview. But in order to be useful, depth interviews do not necessarily ave to be projective in nature; even non-projective depth interviews can reveal important aspects of sycho-social situation for understanding the attitudes of people.

8. **Content-analysis:** Content-analysis consists of analysing the contents of documentary materials such as books, magazines, newspapers and the contents of all other verbal materials which can be either spoken or printed. Content-analysis prior to 1940's was mostly quantitative analysis of documentary materials concerning certain characteristics that can be identified and counted. But since 1950's content-analysis is mostly qualitative analysis concerning the general import or message of the existing documents.

Content-analysis is a central activity whenever one is concerned with the study of the nature of ne verbal materials. A review of research in any area, for instance, involves the analysis of the ontents of research articles that have been published. The analysis may be at a relatively simple evel or may be a subtle one. It is at a simple level when we pursue it on the basis of certain haracteristics of the document or verbal materials that can be identified and counted (such as on the asis of major scientific concepts in a book). It is at a subtle level when researcher makes a study of ne attitude, say of the press towards education by feature writers.

6.4 COLLECTION OF SECONDARY DATA

Secondary data means data that are already available i.e., they refer to the data which have already een collected and analysed by someone else. When the researcher utilises secondary data, then he as to look into various sources from where he can obtain them. In this case he is certainly not onfronted with the problems that are usually associated with the collection of original data. Secondary lata may either be published data or unpublished data. Usually published data are available in: (a) various publications of the central, state are local governments; (b) various publications of foreign governments or of international bodies and their subsidiary organisations; (c) technical and trade ournals; (d) books, magazines and newspapers; (e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (f) reports prepared by research scholars, universities, economists, etc. in different fields; and (g) public records and statistics, historical documents, and other sources of published information.

Nowadays, data are published on the websites also. For example, the official website of Reserve bank of India, National Stock Exchange, etc. Most of the data are freely available and provided in Excel worksheets. The sources of unpublished data are many; they may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/private individuals and organisations.

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study. In this connection Dr. A.L. Bowley very aptly observes that it is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticise arguments that can be based on them.

By way of caution, the researcher, before using secondary data, must see that they poss following characteristics:

1. **Reliability of data:** The reliability can be tested by finding out such things about the s data: (a) Who collected the data? (b) What were the sources of data? (c) Were th collected by using proper methods (d) At what time were they collected? (e) Was th any bias of the compiler? (f) What level of accuracy was desired? (g) Was it achieve
2. **Suitability of data:** The data that are suitable for one enquiry may not necessarily found suitable in another enquiry. Hence, if the available data are found to be unsuitab they should not be used by the researcher. In this context, the researcher must very carefu scrutinise the definition of various terms and units of collection used at the time of collecti the data from the primary source originally. Similarly, the object, scope and nature of original enquiry must also be studied. If the researcher finds differences in these, the da will remain unsuitable for the present enquiry and should not be used.
3. **Adequacy of data:** If the level of accuracy achieved in data is found inadequate for t purpose of the present enquiry, they will be considered as inadequate and should not used by the researcher. The data will also be considered inadequate, if they are related an area which may be either narrower or wider than the area of the present enquiry.

From all this we can say that it is very risky to use the already available data. The alrea available data should be used by the researcher only when he finds them reliable, suitable a adequate. But he should not blindly discard the use of such data if they are readily available fr authentic sources and are also suitable and adequate for in that case it will not be economical spend time and energy in field surveys for collecting information. At times, there may be wealth usable information in the already available data which must be used by an intelligent researcher b with due precaution.

6.5 SELECTION OF APPROPRIATE METHOD FOR DATA COLLECTION

Thus, there are various methods of data collection. As such the researcher must judiciously sele the method/methods for his own study, keeping in view the following factors:

1. **Nature, scope and object of enquiry:** This constitutes the most important factor affecti the choice of a particular method. The method selected should be such that it suits the ty of enquiry that is to be conducted by the researcher. This factor is also important in decidi whether the data already available (secondary data) are to be used or the data not y available (primary data) are to be collected.
2. **Availability of funds:** Availability of funds for the research project determines to a lar extent the method to be used for the collection of data. When funds at the disposal of th researcher are very limited, he will have to select a comparatively cheaper method whic may not be as efficient and effective as some other costly method. Finance, in fact, is a b constraint in practice and the researcher has to act within this limitation.
3. **Time factor:** Availability of time has also to be taken into account in deciding a particul method of data collection. Some methods take relatively more time, whereas with othe the data can be collected in a comparatively shorter duration. The time at the disposal the researcher, thus, affects the selection of the method by which the data are to be collecte

4. **Precision required:** Precision required is yet another important factor to be considered at the time of selecting the method of collection of data.

But one must always remember that each method of data collection has its uses and none is perior in all situations. For instance, telephone interview method may be considered appropriate suming telephone population) if funds are restricted, time is also restricted and the data is to be lected in respect of few items with or without a certain degree of precision. In case funds permit d more information is desired, personal interview method may be said to be relatively better. In se time is ample, funds are limited and much information is to be gathered with no precision, then il-questionnaire method can be regarded more reasonable. When funds are ample, time is also ple and much information with no precision is to be collected, then either personal interview or the il-questionnaire or the joint use of these two methods may be taken as an appropriate method of llecting data. Where a wide geographic area is to be covered, the use of mail-questionnaires pplemented by personal interviews will yield more reliable results per rupee spent than either ethod alone. The secondary data may be used in case the researcher finds them reliable, adequate d appropriate for his research. While studying motivating influences in market researches or studying ople's attitudes in psychological/social surveys, we can resort to the use of one or more of the ojective techniques stated earlier. Such techniques are of immense value in case the reason is tainable from the respondent who knows the reason but does not want to admit it or the reason lates to some underlying psychological attitude and the respondent is not aware of it. But when the spondent knows the reason and can tell the same if asked, than a non-projective questionnaire, ing direct questions, may yield satisfactory results even in case of attitude surveys. Since projective chniques are as yet in an early stage of development and with the validity of many of them remaining open question, it is usually considered better to rely on the straight forward statistical methods ith only supplementary use of projective techniques. Nevertheless, in pre-testing and in searching r hypotheses they can be highly valuable.

Thus, the most desirable approach with regard to the selection of the method depends on e nature of the particular problem and on the time and resources (money and personnel) vailable along with the desired degree of accuracy. But, over and above all this, much depends pon the ability and experience of the researcher. Dr. A.L. Bowley's remark in this context is ery appropriate when he says that "in collection of statistical data common sense is the chief equisite and experience the chief teacher."

6.6 CASE STUDY METHOD

he case study method is a very popular form of qualitative analysis and involves a careful and omplete observation of a social unit, be that unit a person, a family, an institution, a cultural group or ven the entire community. It is a method of study in depth rather than breadth. The case study laces more emphasis on the full analysis of a limited number of events or conditions and their nterrelations. The case study deals with the processes that take place and their interrelationship. hus, case study is essentially an intensive investigation of the particular unit under consideration. he object of the case study method is to locate the factors that account for the behaviour-patterns f the given unit as an integrated totality.

haracteristics: The important characteristics of the case study method are as under:

1. Under this method the researcher can take one single social unit or more of such units for his study purpose; he may even take a situation to study the same comprehensively.

2. Here the selected unit is studied intensively i.e., it is studied in minute details. Generally, t study extends over a long period of time to ascertain the natural history of the unit so as obtain enough information for drawing correct inferences.
3. In the context of this method we make complete study of the social unit covering all face Through this method we try to understand the complex of factors that are operative with a social unit as an integrated totality.
4. Under this method the approach happens to be qualitative and not quantitative. Mo quantitative information is not collected. Every possible effort is made to collect informati concerning all aspects of life. As such, case study deepens our perception and gives u clear insight into life. For instance, under this method we not only study how many crim a man has done but shall peep into the factors that forced him to commit crimes when w are making a case study of a man as a criminal. The objective of the study may be suggest ways to reform the criminal.
5. In respect of the case study method an effort is made to know the mutual inter-relationsh of causal factors.
6. Under case study method the behaviour pattern of the concerning unit is studied direct and not by an indirect and abstract approach.
7. Case study method results in fruitful hypotheses along with the data which may be helpf in testing them, and thus it enables the generalised knowledge to get richer and richer. In i absence, generalised social science may get handicapped.

Evolution and scope: The case study method is a widely used systematic field research techniqu in sociology these days. The credit for introducing this method to the field of social investigation goe to Frederic Le Play who used it as a hand-maiden to statistics in his studies of family budget Herbert Spencer was the first to use case material in his comparative study of different cultures. D William Healy resorted to this method in his study of juvenile delinquency, and considered it as better method over and above the mere use of statistical data. Similarly, anthropologists, historian novelists and dramatists have used this method concerning problems pertaining to their areas o interests. Even management experts use case study methods for getting clues to several managemen problems. In brief, case study method is being used in several disciplines. Not only this, its use i increasing day by day.

Assumptions: The case study method is based on several assumptions. The important assumption may be listed as follows:

(i) The assumption of uniformity in the basic human nature in spite of the fact that huma behaviour may vary according to situations.
(ii) The assumption of studying the natural history of the unit concerned.
(iii) The assumption of comprehensive study of the unit concerned.

Major phases involved: Major phases involved in case study are as follows:

(i) Recognition and determination of the status of the phenomenon to be investigated or th unit of attention.
(ii) Collection of data, examination and history of the given phenomenon.

(iii) Diagnosis and identification of causal factors as a basis for remedial or developmental treatment.

(iv) Application of remedial measures i.e., treatment and therapy (this phase is often characterised as case work).

(v) Follow-up programme to determine effectiveness of the treatment applied.

dvantages: There are several advantages of the case study method that follow from the various aracteristics outlined above. Mention may be made here of the important advantages.

(i) Being an exhaustive study of a social unit, the case study method enables us to understand fully the behaviour pattern of the concerned unit. In the words of Charles Horton Cooley, "case study deepens our perception and gives us a clearer insight into life.... It gets at behaviour directly and not by an indirect and abstract approach."

(ii) Through case study a researcher can obtain a real and enlightened record of personal experiences which would reveal man's inner strivings, tensions and motivations that drive him to action along with the forces that direct him to adopt a certain pattern of behaviour.

(iii) This method enables the researcher to trace out the natural history of the social unit and its relationship with the social factors and the forces involved in its surrounding environment.

(iv) It helps in formulating relevant hypotheses along with the data which may be helpful in testing them. Case studies, thus, enable the generalised knowledge to get richer and richer.

(v) The method facilitates intensive study of social units which is generally not possible if we use either the observation method or the method of collecting information through schedules. This is the reason why case study method is being frequently used, particularly in social researches.

(vi) Information collected under the case study method helps a lot to the researcher in the task of constructing the appropriate questionnaire or schedule for the said task requires thorough knowledge of the concerning universe.

(vii) The researcher can use one or more of the several research methods under the case study method depending upon the prevalent circumstances. In other words, the use of different methods such as depth interviews, questionnaires, documents, study reports of individuals, letters, and the like is possible under case study method.

(viii) Case study method has proved beneficial in determining the nature of units to be studied along with the nature of the universe. This is the reason why at times the case study method is alternatively known as "mode of organising data".

(ix) This method is a means to well understand the past of a social unit because of its emphasis of historical analysis. Besides, it is also a technique to suggest measures for improvement in the context of the present environment of the concerned social units.

(x) Case studies constitute the perfect type of sociological material as they represent a real record of personal experiences which very often escape the attention of most of the skilled researchers using other techniques.

(xi) Case study method enhances the experience of the researcher and this in turn increases his analysing ability and skill.

(xii) This method makes possible the study of social changes. On account of the minute study of the different facets of a social unit, the researcher can well understand the social change then and now. This also facilitates the drawing of inferences and helps in maintaining th continuity of the research process. In fact, it may be considered the gateway to and at th same time the final destination of abstract knowledge.

(xiii) Case study techniques are indispensable for therapeutic and administrative purposes. Th are also of immense value in taking decisions regarding several management problem Case data are quite useful for diagnosis, therapy and other practical case problems.

Limitations: Important limitations of the case study method may as well be highlighted.

(i) Case situations are seldom comparable and as such the information gathered in case studi is often not comparable. Since the subject under case study tells history in his own word logical concepts and units of scientific classification have to be read into it or out of it by th investigator.

(ii) Read Bain does not consider the case data as significant scientific data since they do n provide knowledge of the "impersonal, universal, non-ethical, non-practical, repetitive aspec of phenomena."* Real information is often not collected because the subjectivity of th researcher does enter in the collection of information in a case study.

(iii) The danger of false generalisation is always there in view of the fact that no set rules ar followed in collection of the information and only few units are studied.

(iv) It consumes more time and requires lot of expenditure. More time is needed under cas study method since one studies the natural history cycles of social units and that too minutel

(v) The case data are often vitiated because the subject, according to Read Bain, may writ what he thinks the investigator wants; and the greater the rapport, the more subjective th whole process is.

(vi) Case study method is based on several assumptions which may not be very realistic a times, and as such the usefulness of case data is always subject to doubt.

(vii) Case study method can be used only in a limited sphere, it is not possible to use it in case o a big society. Sampling is also not possible under a case study method.

(viii) Response of the investigator is an important limitation of the case study method. He ofte thinks that he has full knowledge of the unit and can himself answer about it. In case th same is not true, then consequences follow. In fact, this is more the fault of the researche rather than that of the case method.

Despite the above stated limitations, we find that case studies are being undertaken in severa disciplines, particularly in sociology, as a tool of scientific research in view of the several advantage indicated earlier. Most of the limitations can be removed if researchers are always conscious o these and are well trained in the modern methods of collecting case data and in the scientific technique of assembling, classifying and processing the same. Besides, case studies, in modern times, can be conducted in such a manner that the data are amenable to quantification and statistical treatment Possibly, this is also the reason why case studies are becoming popular day by day.

---

* Pauline V. Young, *Scientific social surveys and research*, p. 262.

PROBLEMS

1. Enumerate the different methods of collecting data. Which one is the most suitable for conducting enquiry regarding family welfare programme in India? Explain its merits and demerits.
2. "It is never safe to take published statistics at their face value without knowing their meaning and limitations." Elucidate this statement by enumerating and explaining the various points which you would consider before using any published data. Illustrate your answer by examples wherever possible.
3. Examine the merits and limitations of the observation method in collecting material. Illustrate your answer with suitable examples.
4. Describe some of the major projective techniques and evaluate their significance as tools of scientific social research.
5. How does the case study method differ from the survey method? Analyse the merits and limitations of case study method in sociological research.
6. Clearly explain the difference between collection of data through questionnaires and schedules.
7. Discuss interview as a technique of data collection.
8. Write short notes on:
   (a) Depth interviews;
   (b) Important aspects of a questionnaire;
   (c) Pantry and store audits;
   (d) Thematic Apperception Test;
   (e) Holtzman Inkbolt Test.
9. What are the guiding considerations in the construction of questionnaire? Explain.
10. Critically examine the following:
    (i) Interviews introduce more bias than does the use of questionnaire.
    (ii) Data collection through projective techniques is considered relatively more reliable.
    (iii) In collection of statistical data commonsense is the chief requisite and experience the chief teacher.
11. Distinguish between an experiment and survey. Explain fully the survey method of research.
12. "Experimental method of research is not suitable in management field." Discuss, what are the problems in the introduction of this research design in business organisation?

7

Chapter

Data Preparation

The data, after collection, has to be prepared for analysis. The collected data is raw and it must converted to the form that is suitable for the required analysis. The results of the analysis are affect a lot by the form of the data. So, proper data preparation is a must to get reliable results.

7.1 DATA PREPARATION PROCESS

The plan of data analysis is decided in advance before collecting the data. Data preparation proce is guided by that plan of data analysis. Important steps of data preparation process are as follows

(i) Questionnaire checking
(ii) Editing
(iii) Coding
(iv) Classification
(v) Tabulations
(vi) Graphical representation
(vii) Data cleaning
(viii) Data adjusting

We below describe each of the above processes.

7.1.1 Questionnaire Checking

When the data is collected through questionnaires, the first step of data preparation process is t check the questionnaires if they are acceptable or not. This involves the examination of a questionnaires for their completeness and interviewing quality. Usually this step is undertaken at th time of data collection. If questionnaires checking was not done at the time of collection, it should b done later. A questionnaire may not be acceptable if :

(i) It is incomplete partially or fully.
(ii) It is answered by a person who has inadequate knowledge or does not qualify for the participatio

i) It is answered in such a way which gives the impression that the respondent could not understand the questions.

If sufficient number of questionnaires are not accepted the researcher may like to collect more ta.

1.2 Editing

liting of data is a process of examining the collected raw data (specially in surveys) to detect errors d omissions and to correct these when possible. As a matter of fact, editing involves a careful rutiny of the completed questionnaires and/or schedules. Editing is done to assure that the data are curate, consistent with other facts gathered, uniformly entered, as completed as possible and have een well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and ntral editing. *Field editing* consists in the review of the reporting forms by the investigator for mpleting (translating or rewriting) what the latter has written in abbreviated and/or in illegible form the time of recording the respondents' responses. This type of editing is necessary in view of the ct that individual writing styles often can be difficult for others to decipher. This sort of editing ould be done as soon as possible after the interview, preferably on the very day or on the next day. hile doing field editing, the investigator must restrain himself and must not correct errors of omission y simply guessing what the informant would have said if the question had been asked.

Central editing should take place when all forms or schedules have been completed and returned the office. This type of editing implies that all forms should get a thorough editing by a single editor a small study and by a team of editors in case of a large inquiry. Editor(s) may correct the obvious rrors such as an entry in the wrong place, entry recorded in months when it should have been ecorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes etermine the proper answer by reviewing the other information in the schedule. At times, the espondent can be contacted for clarification. The editor must strike out the answer if the same is appropriate and he has no basis for determining the correct answer or the response. In such a case n editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be ropped from the final results, especially in the context of mail surveys.

Editors must keep in view several points while performing their work: (a) They should be familiar ith instructions given to the interviewers and coders as well as with the editing instructions supplied them for the purpose. (b) While crossing out an original entry for one reason or another, they hould just draw a single line on it so that the same may remain legible. (c) They must make entries if any) on the form in some distinctive colur and that too in a standardised form. (d) They should itial all answers which they change or supply. (e) Editor's initials and the date of editing should be laced on each completed form or schedule.

.1.3 Coding

oding refers to the process of assigning numerals or other symbols to answers so that responses an be put into a limited number of categories or classes. Such classes should be appropriate to the esearch problem under consideration. They must also possess the characteristic of exhaustiveness i.e., there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to observed is that of unidimensionality by which is meant that every class is defined in terms of on one concept.

Coding is necessary for efficient analysis and through it the several replies may be reduced to small number of classes which contain the critical information required for analysis. Coding decisio should usually be taken at the designing stage of the questionnaire. This makes it possible to precod the questionnaire choices and which in turn is helpful for computer tabulation as one can straig forward key punch from the original questionnaires. But in case of hand coding some standa method may be used. One such standard method is to code in the margin with a coloured pencil. Th other method can be to transcribe the data from the questionnaire to a coding sheet. Whateve method is adopted, one should see that coding errors are altogether eliminated or reduced to th minimum level.

7.1.4 Classification

Most research studies result in a large volume of raw data which must be reduced into homogeneou groups if we are to get meaningful relationships. This fact necessitates classification of data whic happens to be the process of arranging data in groups or classes on the basis of common characteristic Data having a common characteristic are placed in one class and in this way the entire data ge divided into a number of groups or classes. Classification can be one of the following two type depending upon the nature of the phenomenon involved:

(a) *Classification according to attributes:* As stated above, data are classified on the basi of common characteristics which can either be descriptive (such as literacy, sex, honesty etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refe to qualitative phenomenon which cannot be measured quantitatively; only their presence o absence in an individual item can be noticed. Data obtained this way on the basis of certai attributes are known as *statistics of attributes* and their classification is said to b classification according to attributes.

Such classification can be simple classification or manifold classification. In simpl classification we consider only one attribute and divide the universe into two classes—on class consisting of items possessing the given attribute and the other class consisting o items which do not possess the given attribute. But in manifold classification we conside two or more attributes simultaneously, and divide that data into a number of classes (tota number of classes of final order is given by 2^n, where n = number of attributes considered). Whenever data are classified according to attributes, the researcher must see that the attributes are defined in such a manner that there is least possibility of any doubt/ambiguity concerning the said attributes.

(b) *Classification according to class-intervals:* Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to income, production, age, weight, etc. come under this

---

[*] Classes of the final order are those classes developed on the basis of 'n' attributes considered. For example, if attribute A and B are studied and their presence is denoted by A and B respectively and absence by a and b respectively, then we hav four classes of final order viz., class AB, class Ab, class aB, and class ab.

category. Such data are known as *statistics of variables* and are classified on the basis of class intervals. For instance, persons whose incomes, say, are within Rs 201 to Rs 400 can form one group, those whose incomes are within Rs 401 to Rs 600 can form another group and so on. In this way the entire data may be divided into a number of groups or classes or what are usually called, 'class-intervals.' Each group of class-interval, thus, has an upper limit as well as a lower limit which are known as class limits. The difference between the two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution. Classification according to class intervals usually involves the following three main problems:

(i) *How may classes should be there? What should be their magnitudes?*

There can be no specific answer with regard to the number of classes. The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class-intervals should be of equal magnitudes, but in some cases unequal magnitudes may result in better classification. Hence the researcher's objective judgement plays an important part in this connection. Multiples of 2, 5 and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by H.A. Sturges, determining the size of class interval:

$$\boxed{i = R/(1 + 3.3 \log N)}$$

where

i = size of class interval;

R = Range (i.e., difference between the values of the largest item and smallest item among the given items);

N = Number of items to be grouped.

It should also be kept in mind that in case one or two or very few items have very high or very ow values, one may use what are known as open-ended intervals in the overall frequency distribution. Such intervals may be expressed like under Rs. 500 or Rs. 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researcher must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are to be classified.

(ii) *How to choose class limits?*

While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by 2) of a class-interval and the actual average of items of that class interval should remain as close to each other as possible.

Consistent with this, the class limits should be located at multiples of 2, 5, 10, 20, 1 and such other figures. Class limits may generally be stated in any of the followi forms:

Exclusive type class intervals: They are usually stated as follows:

10–20
20–30
30–40
40–50

The above intervals should be read as under:

10 and under 20
20 and under 30
30 and under 40
40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to th upper limit of a class are grouped in the next higher class. For example, an item who value is exactly 30 would be put in 30–40 class interval and not in 20–30 class interva In simple words, we can say that under exclusive type class intervals, the upper limit a class interval is excluded and items with values less than the upper limit (but not le than the lower limit) are put in the given class interval.

Inclusive type class intervals: They are usually stated as follows:

11–20
21–30
31–40
41–50

In inclusive type class intervals the upper limit of a class interval is also included in th concerning class interval. Thus, an item whose value is 20 will be put in 11–20 cla interval. The stated upper limit of the class interval 11–20 is 20 but the real limit 20.99999 and as such 11–20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be a discrete one (i.e., can be measure and stated only in integers), then we should adopt inclusive type classification. But when th phenomenon happens to be a continuous one capable of being measured in fractions as well, we ca use exclusive type class intervals.*

(iii) *How to determine the frequency of each class?*

This can be done either by tally sheets or by mechanical aids. Under the technique c tally sheet, the class-groups are written on a sheet of paper (commonly known as th tally sheet) and for each item a stroke (usually a small vertical line) is marked agains the class group in which it falls. The general practice is that after every four smal vertical lines in a class group, the fifth line for the item falling in the same group, i indicated as horizontal line through the said four lines and the resulting flower (IIII) represents five items. All this facilitates the counting of items in each one of the class groups. An illustrative tally sheet can be shown as under:

---

* The stated limits of class intervals are different than true limits. We should use true or real limits keeping in view th nature of the given phenomenon.

Table 7.1: An Illustrative Tally Sheet for Determining the Number of 70 Families in Different Income Groups

Income groups (Rupees)	*Tally mark*	*Number of families or (Class frequency)*
Below 400	~~IIII~~ ~~IIII~~ III	13
401–800	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
801–1200	~~IIII~~ ~~IIII~~ II	12
1201–1600	~~IIII~~ ~~IIII~~ ~~IIII~~ III	18
1601 and above	~~IIII~~ II	7
Total		70

Alternatively, class frequencies can be determined, specially in case of large inquires and surveys, mechanical aids i.e., with the help of machines viz., sorting machines that are available for the irpose. Some machines are hand operated, whereas other work with electricity. There are machines hich can sort out cards at a speed of something like 25000 cards per hour. This method is fast but pensive.

.1.5 Tabulation

hen a mass of data has been assembled, it becomes necessary for the researcher to arrange the ame in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, bulation is the process of summarising raw data and displaying the same in compact form (i.e., in e form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement f data in columns and rows.

Tabulation is essential because of the following reasons.

1. It conserves space and reduces explanatory and descriptive statement to a minimum.
2. It facilitates the process of comparison.
3. It facilitates the summation of items and the detection of errors and omissions.
4. It provides a basis for various statistical computations.

Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on e size and type of study, cost considerations, time pressures and the availability of tabulating machines r computers. In relatively large inquiries, we may use mechanical or computer tabulation if other ictors are favourable and necessary facilities are available. Hand tabulation is usually preferred in ase of small inquiries where the number of questionnaires is small and they are of relatively short length. Hand tabulation may be done using the direct tally, the list and tally or the card sort and co methods. When there are simple codes, it is feasible to tally directly from the questionnaire. Un this method, the codes are written on a sheet of paper, called tally sheet, and for each respons stroke is marked against the code in which it falls. Usually after every four strokes against a particu code, the fifth response is indicated by drawing a diagonal or horizontal line through the strok These groups of five are easy to count and the data are sorted against each code conveniently. In listing method, the code responses may be transcribed onto a large work-sheet, allowing a line each questionnaire. This way a large number of questionnaires can be listed on one work sh Tallies are then made for each question. The card sorting method is the most flexible hand tabulati In this method the data are recorded on special cards of convenient size and shape with a series holes. Each hole stands for a code and when cards are stacked, a needle passes through particu hole representing a particular code. These cards are then separated and counted. In this w frequencies of various codes can be found out by the repetition of this technique. We can as well the mechanical devices or the computer facility for tabulation purpose in case we want quick resu our budget permits their use and we have a large volume of straight forward tabulation involvin number of cross-breaks.

Tabulation may also be classified as simple and complex tabulation. The former type of tabulat gives information about one or more groups of independent questions, whereas the latter type tabulation shows the division of data in two or more categories and as such is deigned to g information concerning one or more sets of inter-related questions. Simple tabulation generally resu in one-way tables which supply answers to questions about one characteristic of data only. against this, complex tabulation usually results in two-way tables (which give information about t inter-related characteristics of data), three-way tables (giving information about three interrela characteristics of data) or still higher order tables, also known as manifold tables, which sup information about several interrelated characteristics of data. Two-way tables, three-way tables manifold tables are all examples of what is sometimes described as cross tabulation.

Generally accepted principles of tabulation: Such principles of tabulation, particularly constructing statistical tables, can be briefly states as follows:*

1. Every table should have a clear, concise and adequate title so as to make the table intelligi without reference to the text and this title should always be placed just above the body the table.
2. Every table should be given a distinct number to facilitate easy reference.
3. The column headings (captions) and the row headings (stubs) of the table should be cl and brief.
4. The units of measurement under each heading or sub-heading must always be indicate
5. Explanatory footnotes, if any, concerning the table should be placed directly beneath t table, along with the reference symbols used in the table.
6. Source or sources from where the data in the table have been obtained must be indicat just below the table.
7. Usually the columns are separated from one another by lines which make the table mo readable and attractive. Lines are always drawn at the top and bottom of the table a below the captions.

---

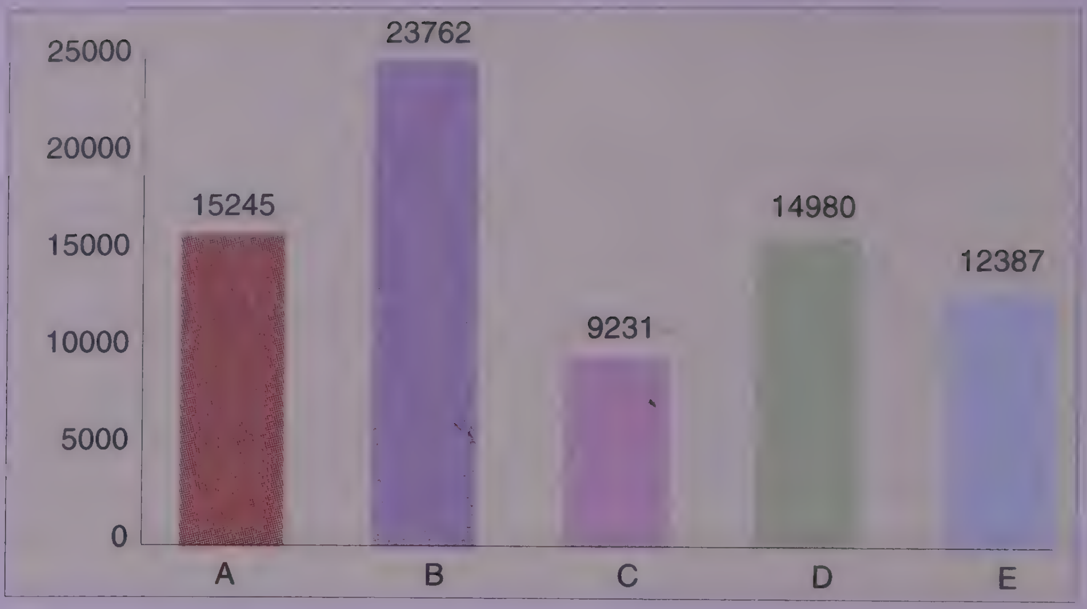
* All these points constitute the characteristics of a good table.

8. There should be thick lines to separate the data under one class from the data under another class and the lines separating the sub-divisions of the classes should be comparatively thin lines.
9. The columns may be numbered to facilitate reference.
10. Those columns whose data are to be compared should be kept side by side. Similarly, percentages and/or averages must also be kept close to the data.
11. It is generally considered better to approximate figures before tabulation as the same would reduce unnecessary details in the table itself.
12. In order to emphasis the relative significance of certain categories, different kinds of type, spacing and indentations may be used.
13. It is important that all column figures be properly aligned. Decimal points and (+) or (–) signs should be in perfect alignment.
14. Abbreviations should be avoided to the extent possible and ditto marks should not be used in the table.
15. Miscellaneous and exceptional items, if any, should be usually placed in the last row of the table.
16. Table should be made as logical, clear, accurate and simple as possible. If the data happen to be very large, they should not be crowded in a single table for that would make the table unwieldy and inconvenient.
17. Total of rows should normally be placed in the extreme right column and that of columns should be placed at the bottom.
18. The arrangement of the categories in a table may be chronological, geographical, alphabetical or according to magnitude to facilitate comparison. Above all, the table must suit the needs and requirements of an investigation.

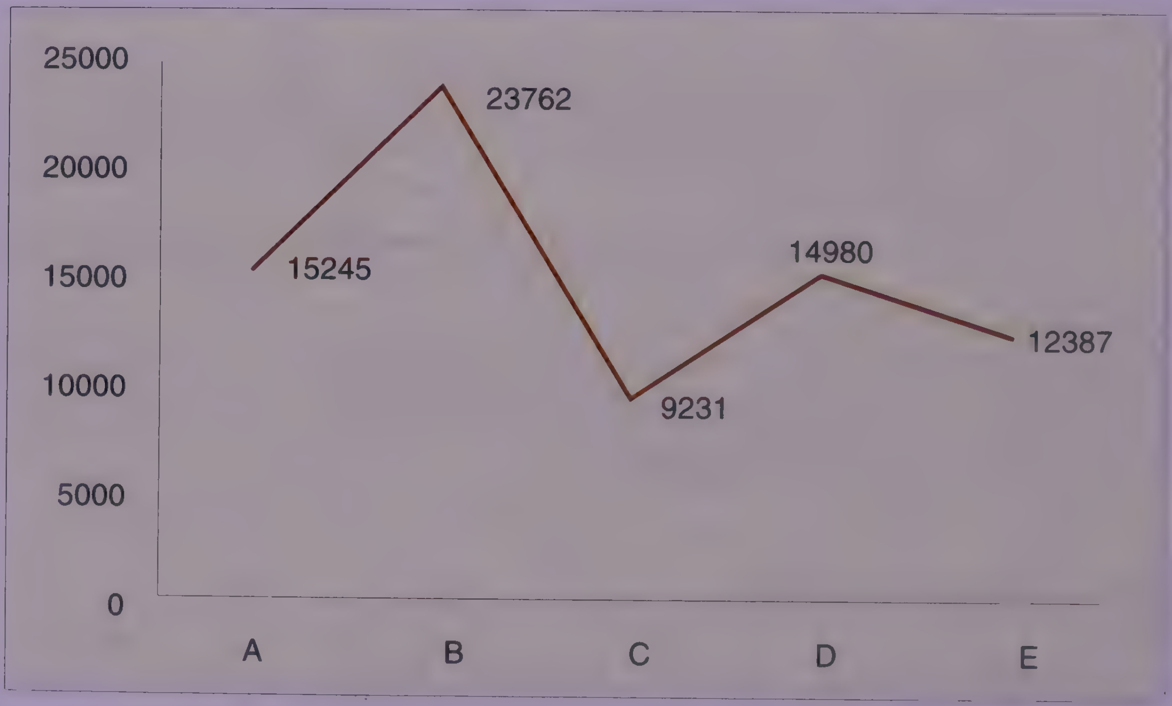
1.6 Graphical Representation

aphs help to understand the data easily. All statistical packages, MS Excel, and OpenOffice.org fer a wide range of graphs. In case of qualitative data (or categorized data), most common graphs e bar charts and pie charts.

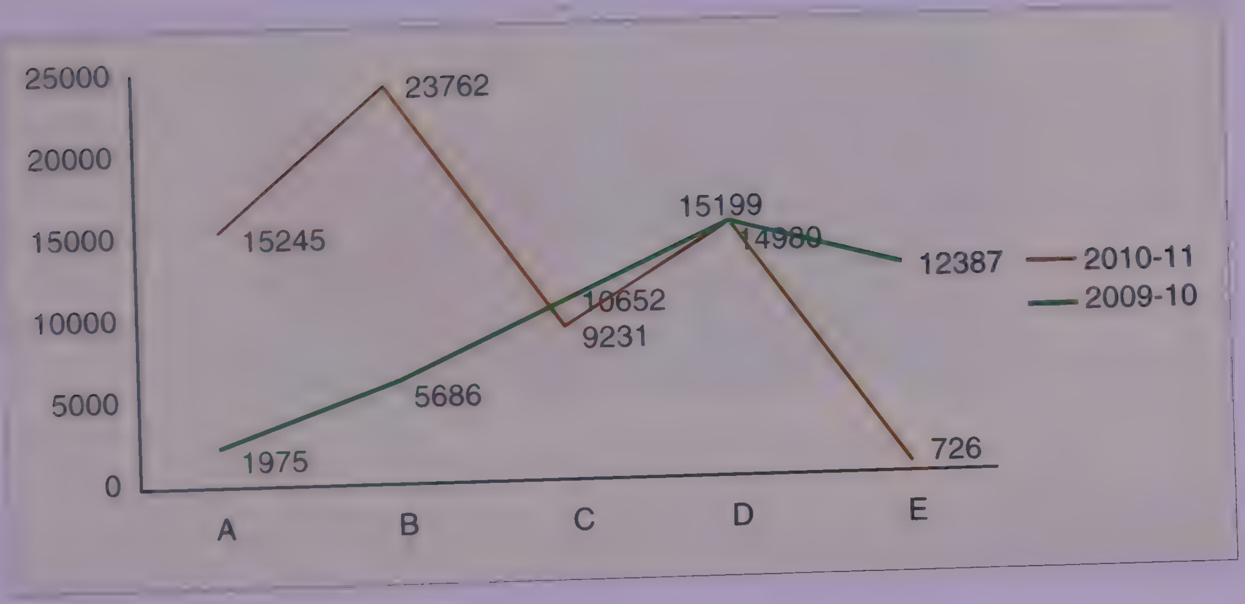
Bar chart: A bar chart consists of a series of rectangles (or bars). The height of each rectangle determined by the frequency of that category. Suppose that the sales of a popular soft drink in the ar 2010-11, in five geographical regions, denoted as A, B, C, D, and E, are 15245, 23762, 9231, 980, and 12387, respectively, measured in 10,000 USD. A bar chart of this data is as below.



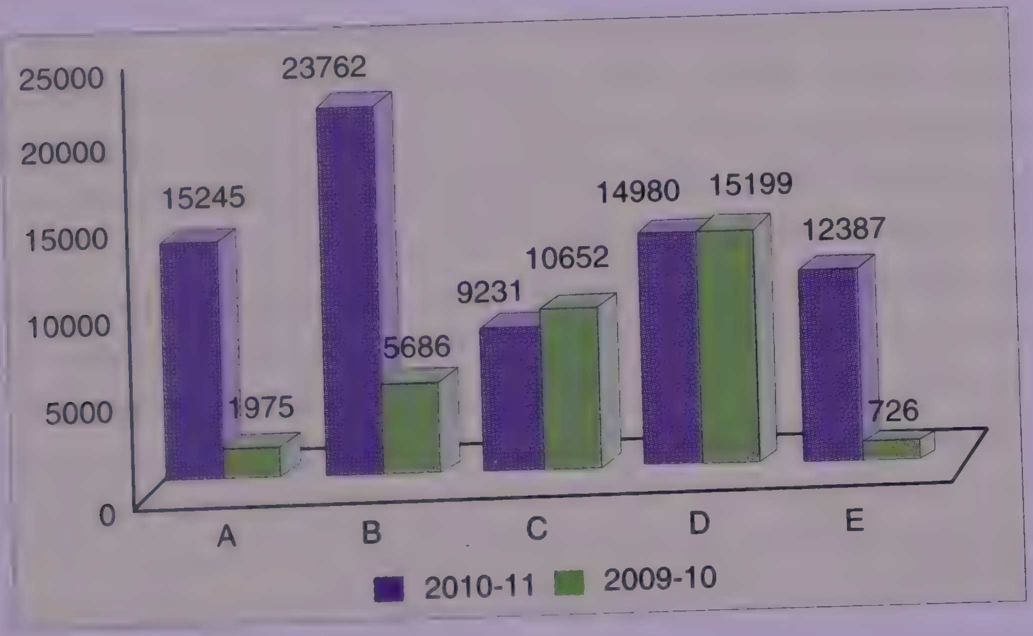
A line chart can also be plotted in this data by connecting the midpoints of each rectangle below:



Line charts are useful when we wish to compare to data sets as we can overlap to line charts. F example, the sales data of the same soft drink in the same geographical regions in the year 2009- were 1975, 5686, 10652, 15199 and 726, respectively, measured in 10,000 USD. The line cha showing the data for both the years is

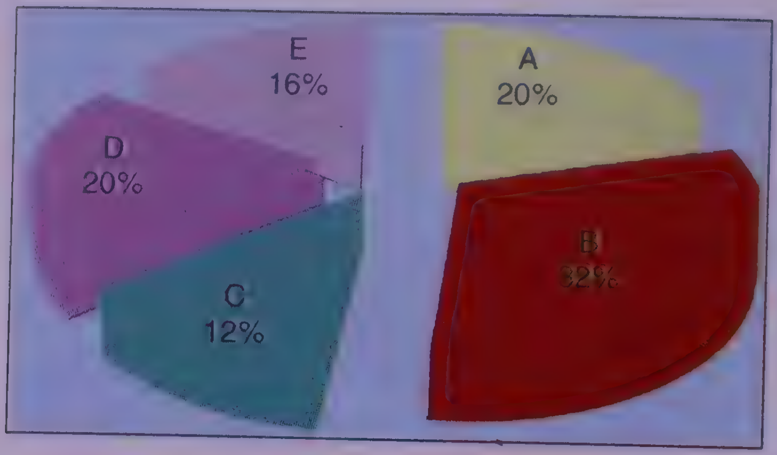


nother graphical representation of the same data is using multiple bars as below:



Pie chart: A pie chart is used to emphasize relative proportion or shares of each category. It's a ircular chart divided into sectors, illustrating relative frequencies. The relative frequency in each ategory or sector is proportional to the arc length of that sector or the area of that sector or the entral angle of that sector. Suppose in the previous example, if the soft drink has its market only in he five geographical regions, denoted as A, B, C, D, and E. For the year 2010-11, the sales data in hese regions are 15245, 23762, 9231, 14980, and 12387, respectively, measured in 10,000 USD. A otal sale of the soft drink is 75605 times 10,000 USD. A pie chart can be plotted to have the idea of he shares of different markets.

In case of quantitative data, one important chart is histogram which is a generalization of bar hart. The data is first summarized in terms of class intervals and each bar represents a class interval. The width of the bar is proportional to the width of corresponding class interval. The



area of the bar is proportional to the frequency of corresponding class interval. After making th class intervals in a quantitative data set, a pie chart can also be used to read the share of each cla interval.

7.1.7 Data Cleaning

This includes checking the data for consistency and treatment for missing value. Preliminar consistency checks are made in editing. Here we check the consistency in an extensive manner.

Consistency checks look for the data which are not consistent or outlines. Such data may eithe be discarded or replaced by the mean value. However, the researcher should be careful while doin this. Extreme values or outlines are not always erroneous.

Missing values are the values which are unknown or not answered by the respondent. In plac of such missing values, some neutral value may be used. This neutral value may be the mean o available values. The other option could be to use the pattern of responses to other questions t calculate a suitable substitute to the missing values.

7.1.8 Data Adjusting

Data adjusting is not always necessary but it may improve the quality of analysis sometimes. Thi consists of following methods.

(i) **Weight-assigning:** Each respondent or case is assigned a weight to reflect its importance relative to other respondents or cases.

Using this method, the collected sample can be made a stronger representative of a target population on specific characteristics. For example, the cases of educated people could be assigned higher weights and of uneducated people could be assigned lower weights in some survey. The value 1.0 means unweightage case.

(ii) **Variable Respecification:** This involves creating new variables or modifying existing variables. For example, if the usefulness of a certain product is measured on 10 point scale, it may be reduced on a 4 point scale — 'very useful', 'useful', 'neutral', 'not useful'. Ratio of two variables may also be taken to create a new variable.

Method of dummy variables for respecifying categorical variables is also very popular. Dummy variable is a variable which usually takes numerical values based on the corresponding category in the original variable. For example, a group of people is divided into smokers and non-smokers. We can define a dummy variable taking the value '1' for smokers and '0' for non-smokers.

(iii) Scale Transformation: Scale transformation is done to ensure the comparability with other scales or to make the data suitable for analysis. Different type of characteristics are measured on different scales. For example, attitude variables are measured on continuous scale, life style variables are usually measured on a 5 point Likert scale. So the variables which are measured on different scales, cannot be compared. A common transformation is subtracting all the values of a characteristic by corresponding mean and dividing by corresponding standard deviation.

7.2 SOME PROBLEMS IN PREPARATION PROCESS

Ve can take up the following two problems of processing the data for analytical purposes:

ı) *The problem concerning "Don't know" (or DK) responses:* While processing the data, the esearcher often comes across some responses that are difficult to handle. One category of such esponses may be 'Don't Know Response' or simply DK response. When the DK response group is mall, it is of little significance. But when it is relatively big, it becomes a matter of major concern in vhich case the question arises: Is the question which elicited DK response useless? The answer lepends on two points viz., the respondent actually may not know the answer or the researcher may ail in obtaining the appropriate information. In the first case the concerned question is said to be lright and DK response is taken as legitimate DK response. But in the second case, DK response s more likely to be a failure of the questioning process.

How DK responses are to be dealt with by researchers? The best way is to design better type of questions. Good rapport of interviewers with respondents will result in minimising DK responses. But what about the DK responses that have already taken place? One way to tackle this issue is to estimate the allocation of DK answers from other data in the questionnaire. The other way is to keep DK responses as a separate category in tabulation where we can consider it as a separate reply category if DK responses happen to be legitimate, otherwise we should let the reader make his own decision. Yet another way is to assume that DK responses occur more or less randomly and as such we may distribute them among the other answers in the ratio in which the latter have occurred. Similar results will be achieved if all DK replies are excluded from tabulation and that too without inflating the actual number of other responses.

(b) *Use of percentages:* Percentages are often used in data presentation for they simplify numbers, reducing all of them to a 0 to 100 range. Through the use of percentages, the data are reduced in the standard form with base equal to 100 which fact facilitates relative comparisons. While using percentages, the following rules should be kept in view by researchers:

1. Two or more percentages must not be averaged unless each is weighted by the group size from which it has been derived.
2. Use of too large percentages should be avoided, since a large percentage is difficult to understand and tends to confuse, defeating the very purpose for which percentages are used.
3. Percentages hide the base from which they have been computed. If this is not kept in view, the real differences may not be correctly read.

4. Percentage decreases can never exceed 100 per cent and as such for calculating percentage of decrease, the higher figure should invariably be taken as the base.
5. Percentages should generally be worked out in the direction of the causal-factor in case two-dimension tables and for this purpose we must select the more significant factor ou the two given factors as the causal factor.

7.3 MISSING VALUES AND OUTLIERS

Missing values are the observations which the researcher plan to collect but could not collect or l due to some reason. Many statistical tools cannot be employed when the data set has one or mo missing values. In data collection through asking questions 'Don't know' response may also cre the problem of missing values. Utmost care should be taken by the researcher to avoid the missi values in the data set. Most common methods to deal with the problem of missing value wh conducting the analysis is either to leave the observation, if possible, or to replace the missing val by the arithmetic mean of other collected observation.

Outliers are the observations which are quite different to other observations in the data se Although all the statistical techniques can be employed when data set has outliers, their interpretatio may be misleading. The most common reason of outliers being present in the data set is the recordi error. This error should be corrected while editing and cleaning the data. Consider an example survey of 100 customers in a mall.

If few bulk customers purchasing very large amounts are among the 100 surveyed customers. this survey having outliers (bulk customers) may not be posing any error as bulk customers a always there in the mall along with small customers. However, in a similar survey at a near grocerry shop on a day when there is strike in the mall may include some bulk customers which coul be misleading. Thus outliers should not be ignored as they might have some relevant information pose to a serious risk.

Before detecting the outliers, we need to define them first. Commonly, an observation with value that is more than 3 standard deviations from the mean is considered as an outlier. A scatter pl (discussed later) can also be helpful in identifying the outliers. After identifying an outlier, the researche has to decide what to do with it. The researcher may like to delete it or modify the value of it or retai it as it is. It depends on the knowledge about the cause of that outlier.

7.4 TYPES OF ANALYSIS

As stated earlier, by analysis we mean the computation of certain indices or measures along wit searching for patterns of relationship that exist among the data groups. Analysis, particularly in case of survey or experimental data, involves estimating the values of unknown parameters of the populatio and testing of hypotheses for drawing inferences. Analysis may, therefore, be categorised as descriptive analysis and inferential analysis (Inferential analysis is often known as statistical analysis). Descriptive analysis is largely the study of the distributions of one or more variables involved in the study. In this context we work out various measures that show the size and shape of a distribution(s) along with the study of measuring relationships between two or more variables.

We may as well talk of correlation analysis and causal analysis. *Correlation analysis* studies the joint variation of two or more variables for determining the amount of correlation between two or more variables. *Causal analysis* is concerned with the study of how one or more variables affect anges in another variable. It is thus a study of functional relationships existing between two or re variables. This analysis can be termed as regression analysis. Causal analysis is considered atively more important in experimental researches, whereas in most social and business researches interest lies in understanding and controlling relationships between variables then with determining uses *per se* and as such we consider correlation analysis as relatively more important.

In modern times, with the availability of computer facilities, there has been a rapid development *multivariate analysis.* Usually the following analyses are involved when we make a reference of ıltivariate analysis:

(a) *Multiple regression analysis:* This analysis is adopted when the researcher has one dependent variable which is presumed to be a function of two or more independent variables. The objective of this analysis is to make a prediction about the dependent variable based on its covariance with all the concerned independent variables.

(b) *Multiple discriminant analysis:* This analysis is appropriate when the researcher has a single dependent variable that cannot be measured, but can be classified into two or more groups on the basis of some attribute. The object of this analysis happens to be to predict an entity's possibility of belonging to a particular group based on several predictor variables.

(c) *Multivariate analysis of variance (or multi-ANOVA):* This analysis is an extension of two-way ANOVA, wherein the ratio of among group variance to within group variance is worked out on a set of variables.

(d) *Canonical analysis:* This analysis can be used in case of both measurable and non-measurable variables for the purpose of simultaneously predicting a set of dependent variables from their joint covariance with a set of independent variables.

Inferential analysis is concerned with the various tests of significance for testing hypotheses in der to determine with what validity data can be said to indicate some conclusion or conclusions. It also concerned with the estimation of population values. It is mainly on the basis of inferential ıalysis that the task of interpretation (i.e., the task of drawing inferences and conclusions) is erformed.

7.5 STATISTICS IN RESEARCH

he role of statistics in research is to function as a tool in designing research, analysing its data and rawing conclusions therefrom. Most research studies result in a large volume of raw data which ıust be suitably reduced so that the same can be read easily and for further analysis. Clearly the cience of statistics cannot be ignored by any research worker, even though he may not have occasion use statistical methods in all their details and ramifications. Classification and tabulation, as stated arlier, achieve this objective to some extent, but we have to go a step further and develop certain ndices or measures to summarise the collected/classified data. Only after this we can adopt the rocess of generalisation from small groups (i.e., samples) to population. If fact, there are two major reas of statistics viz., descriptive statistics and inferential statistics. *Descriptive statistics* concern e development of certain indices from the raw data, whereas inferential statistics concern with the rocess of generalisation. *Inferential statistics* are also known as sampling statistics and are mainly oncerned with two major type of problems: (i) the estimation of population parameters, and (ii) the esting of statistical hypotheses.

The important statistical measures that are used to summarise the survey/research data ar (1) measures of central tendency or statistical averages; (2) measures of dispersion; (3) meas of asymmetry (skewness); (4) measures of relationship; and (5) other measures.

Amongst the measures of central tendency, the three most important ones are the arithm average or mean, median and mode. Geometric mean and harmonic mean are also sometimes u

From among the measures of dispersion, variance, and its square root—the standard devia are the most often used measures. Other measures such as mean deviation, range, etc. are used. For comparison purpose, we use mostly the coefficient of standard deviation or the coeffic of variation.

In respect of the measures of skewness and kurtosis, we mostly use the first measure of skew based on mean and mode or on mean and median. Other measures of skewness, based on quart or on the methods of moments, are also used sometimes. Kurtosis is also used to measure peakedness of the curve of the frequency distribution.

Amongst the measures of relationship, Karl Pearson's coefficient of correlation is the freque used measure in case of statistics of variables, whereas Yule's coefficient of association is use case of statistics of attributes. Multiple correlation coefficient, partial correlation coefficient, regress analysis, etc., are other important measures often used by a researcher.

Index numbers, analysis of time series, coefficient of contingency, etc., are other measures t may as well be used by a researcher, depending upon the nature of the problem under study.

We give below a brief outline of some important measures (our of the above listed measur often used in the context of research studies.

PROBLEMS

1. "Processing of data implies editing, coding, classification and tabulation". Describe in brief these fou operations pointing out the significance of each in context of research study.
2. Classification according to class intervals involves three main problems viz., h\How many classe should be there? How to choose class limits? How to determine class frequency? State how thes problems should be tackled by a researcher.
3. Why tabulation is considered essential in a research study? Narrate the characteristics of a good tabl
4. (a) How the problem of DK responses should be dealt with by a researcher? Explain.
   (b) What points one should observe while using percentages in research studies?
5. Write short notes on the followings:
   (i) Data adjusting
   (ii) Data cleaning
   (iii) Questionnaire checking
6. Why scale transformations are made? Explain.
7. How will you treat the missing data?
8. What are the reasons of weight assigning?
9. What are dummy variables? Give example.

escriptive Statistics

8

Chapter

ere are some important measures which help to know the data better. These measures give the a of over all distribution of the observations in the data-set. These measures together can be led as descriptive statistics. In the first section of this chapter we discuss the measures of central dency; in second section, measures of dispersion; measures of asymmetry in third section and asures of relationship in fourth section.

.1 MEASURES OF CENTRAL TENDENCY

ese measures are also called as statistical averages or averages. A measure of central tendency a value around which all the observations have a tendency to cluster. Such a value is considered as most representative figure of the entire data-set. Three most popular and important measures of ntral tendency are mean, median, and mode.

1.1 Mean

is the most common measure of central tendency and may be defined as the value which we get dividing the total of the values of various given items in a series by the total number of items. we n work it out as under:

$$\boxed{\text{Mean (or } \bar{X}) = \frac{\Sigma X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}}$$

nere, $\bar{X}$ = The symbol we use for mean (pronounced as X bar)

Σ = Symbol for summation

X_i = Value of the ith item X, $i = 1, 2, \ldots, n$

n = Total number of items

In case of frequency distribution when the whole data set of size n is summarized in k class tervals, mean is calculated as

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{f_1 X_1 + f_2 X_2 + ... + f_k X_k}{f_1 + f_2 + ... + f_k}$$

where, X_i = Mid-point of *i*th class interval and $f_1 + f_2 + ... + f_k = n$.

Sometimes, instead of calculating the simple mean, as stated above, we may workout the weigh mean for a realistic average. The weighted mean can bè worked out as follows:

$$\overline{X}_w = \frac{\Sigma w_i X_i}{\Sigma w_i}$$

where, $\overline{X}_w$ = Weighted item

w_i = Weight of *i*th item *X*, assigned by the researcher using the some prior knowledge

X_i = Value of the *i*th item *X*

Mean is the simplest and most widely used measure of central tendency. People use mean daily life so much that it has become a synonym of average. For example, Ravi consumes 4 cigarett on an average daily; Suresh drinks approximately 2 litres of water, on an average daily. However, t mean suffers from some limitations. When the data-set has one or more extreme values, the magnitu of mean is affected and it provides a wrong impression of the other values in the data-set, when us to represent the whole data-set.

8.1.2 Median

When the data-set has outliers, mean becomes flowed as a representative of the data-set. In such case, median is used as a measure of central tendency. Median divides the data-set into two equ parts. Half of the items are less than the median and remaining half of the items are larger than th median. In order to obtain the median, we first arrange the data-set into ascending or descendi order. If number of observations in the data set is *n*, then the

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th observation, when } n \text{ is odd;}$$

$$= \frac{1}{2}\left[\left(\frac{n}{2}\right)\text{th observation} + \left(\frac{n}{2}+1\right)\text{th observation}\right], \text{ when } n \text{ is even}$$

For example, median of the data-set 2, 5, 6, 11, 59 is 6, while median of the data-set 5, 6, 11, 5 is $\frac{1}{2}[6+11]$ or 8.5.

Median is a positional average and is used only in the context of qualitative phenomena, fo example, in estimating intelligence, etc., which are often encountered in sociological fields. Median i not useful where items need to be assigned relative importance and weights. It is not frequently use in sampling statistics.

.3 Mode

e most frequently occuring observation in the data-set is mode. Mode is a french word having the aning fashion. It is particularly useful in the study of popular sizes. For example, a manufacturer shoes is usually interested in finding out the size most in demand so that he may manufacture a ger quantity of that size. Like median, mode is also a positional average and is not affected by reme values. Mode is not amenable to algebraic treatment. A data-set may not have any mode or re may be more than one modes in a data-set.

1.4 Other Averages

ometric mean is also useful under certain conditions. It is defined as the *n*th root of the product of values of *n* times in a given series. Symbolically, we can put it thus:

$$\text{Geometric mean (or G.M.)} = \sqrt[n]{\pi X_i}$$
$$= \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \ldots X_n}$$

ere, G.M. = Geometric mean

n = Number of items

X_i = ith value of the variable X

π = Conventional product notation

For instance, the geometric mean of the numbers, 4, 6, and 9 is worked out as

$$\text{G.M.} = \sqrt[3]{4 \cdot 6 \cdot 9}$$
$$= 6$$

The most frequently used application of this average is in the determination of average per cent change i.e., it is often used in the preparation of index numbers or when we deal in ratios.

Harmonic mean is defined as the reciprocal of the average of reciprocals of the values of items a series. Symbolically, we can express it as under:

$$\text{Harmonic mean (H.M.)} = \text{Rec.} \frac{\Sigma\, \text{Rec} X_i}{n}$$
$$= \text{Rec.} \frac{\text{Rec.}\, X_1 + \text{Rec.}\, X_2 + \ldots + \text{Rec.}\, X_n}{n}$$

here, H.M. = Harmonic mean

Rec. = Reciprocal

X_i = ith value of the variable X

n = Number of items.

For instance, the harmonic mean of the numbers 4, 5, and 10 is worked out as

$$\text{H.M.} = \text{Rec.}\frac{1/4+1/5+1/10}{3} = \text{Rec.}\frac{\dfrac{15+12+6}{60}}{3}$$

$$= \text{Rec.}\left(\frac{33}{60}\times\frac{1}{3}\right) = \frac{60}{11} = 5.45$$

Harmonic mean is of limited application, particularly in cases where time and rate are invol The harmonic mean gives largest weight to the smallest item and smallest weight to the largest i As such it is used in cases like time and motion study where time is variable and distance const

The selection of an average mostly depends on the nature, type of objectives of the rese study. One particular type of average cannot be taken as appropriate for all types of studies. chief characteristics and the limitations of the various averages must be kept in view; discrimi use of average is very essential for sound statistical analysis.

8.2 MEASURES OF DISPERSION

An average can represent a data-set only as best as a single value can, but it certainly cannot rev the entire story of the data-set. For example, consider the following data-sets:

(i) 5, 5, 5, 5, 5; (mean = 5, median = 5)
(ii) 3, 4, 5, 6, 7; (mean = 5, median = 5)
(iii) 1, 3, 5, 7, 9; (mean = 5, median = 5).

In all three data-sets above, the average values (mean and median) are the same. What observe here is, an average fails to give any idea about the spread or scatteredness of the observati in the data-set. This scatteredness is also called dispersion. Some important measures of dispers are range, mean deviation, and standard deviation.

8.2.1 Range

Range is the simplest possible measure of dispersion and is defined as the difference between values of the extreme items of a series. Thus,

$$\boxed{\text{Range} = \begin{pmatrix}\text{Highest value of an}\\ \text{item in a series}\end{pmatrix} - \begin{pmatrix}\text{Lowest value of an}\\ \text{item in a series}\end{pmatrix}}$$

The utility of range is that it gives an idea of the variability very quickly. but the drawback is t range is affected very greatly by fluctuations of sampling. Its value is never stable, being based only two values of the variable. As such, range is mostly used as a rough measure of variability a is not considered as an appropriate measure in serious research studies.

8.2.2 Mean Deviation

Mean deviation is the average of difference of the values of items from some average of the seri Such a difference is technically described as deviation. In calculating mean deviation we ignore t us sign of deviations while taking their total for obtaining the mean deviation. Mean deviation is, , obtained as under:

Mean deviation from mean $(\delta_{\bar{X}}) = \dfrac{\Sigma|X_i - \bar{X}|}{n}$, if deviations $|X_i - \bar{X}|$ are obtained from arithmetic average.

or

Mean deviation from median $(\delta_m) = \dfrac{\Sigma|X_i - M|}{n}$, if deviations $|X_i - M|$ are obtained from median.

or

Mean deviation from mode $(\delta_z) = \dfrac{\Sigma|X_i - Z|}{n}$, if deviations $|X_i - Z|$ are obtained from mode.

ere, δ = Symbol for mean deviation (pronounced as delta)
X_i = ith values of the variable X
n = Number of items
$\bar{X}$ = Arithmetic average
M = Median
Z = Mode.

When mean deviation is divided by the average used in finding out the mean deviation itself, the ulting quantity is described as the *coefficient of mean deviation*. Coefficient of mean deviation relative measure of dispersion and is comparable to similar measure of other series. Mean iation and its coefficient are used in statistical studies for judging the variability, and thereby der the study of central tendency of a series more precise by throwing light on the typicalness of average. It is a better measure of variability than range as it takes into consideration the values of items of a series. Even then it is not a frequently used measure as it is not amenable to algebraic cess.

2.3 Standard Deviation

ndard deviation is most widely used measure of dispersion of a series and is commonly denoted the symbol ' σ ' (pronounced as sigma). Standard deviation is defined as the square-root of the erage of squares of deviations, when such deviations for the values of individual items in a series obtained from the arithmetic average. It is worked out as under:

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}}$$

Or

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma f_i (X_i - \bar{X})^2}{\Sigma f_i}}$$

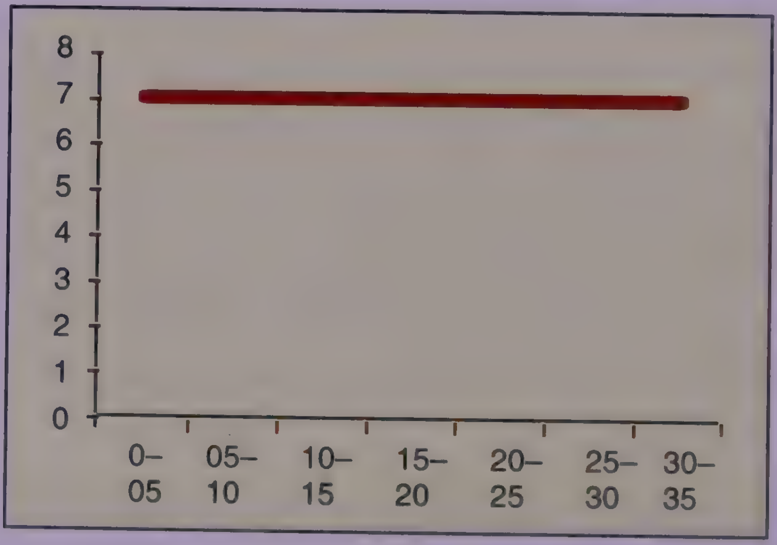
in case of frequency distribution where f_i means the frequency of the ith item.

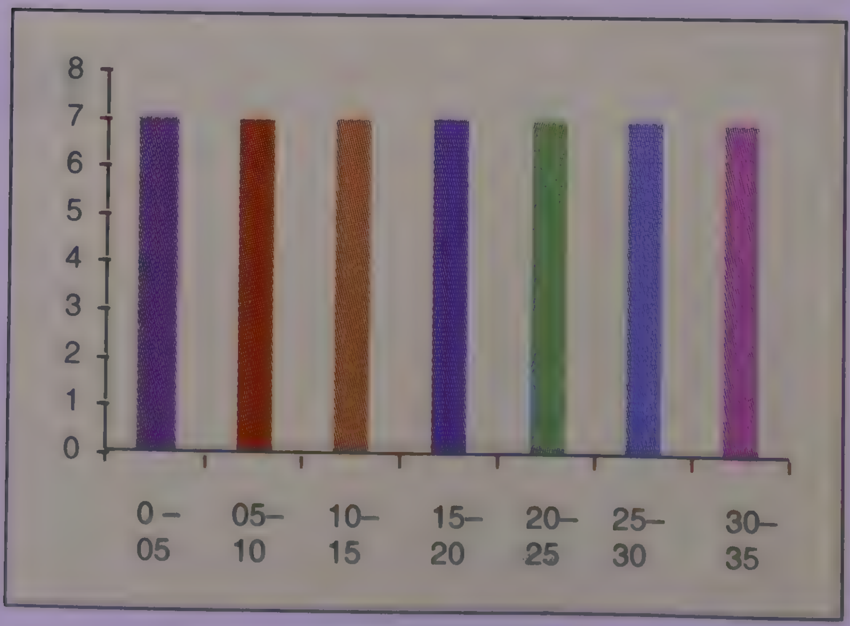
When we divide the standard deviation by the arithmetic average of the series, the resul quantity is known as *coefficient of standard deviation which* happens to be a relative measure . is often used for comparing with similar measure of other series. When this coefficient of stand deviation is multiplied by 100, the resulting figure is known as *coefficient of variation*. Sometim we work out the square of standard deviation, known as *variance*, which is frequently used in context of analysis of variation.

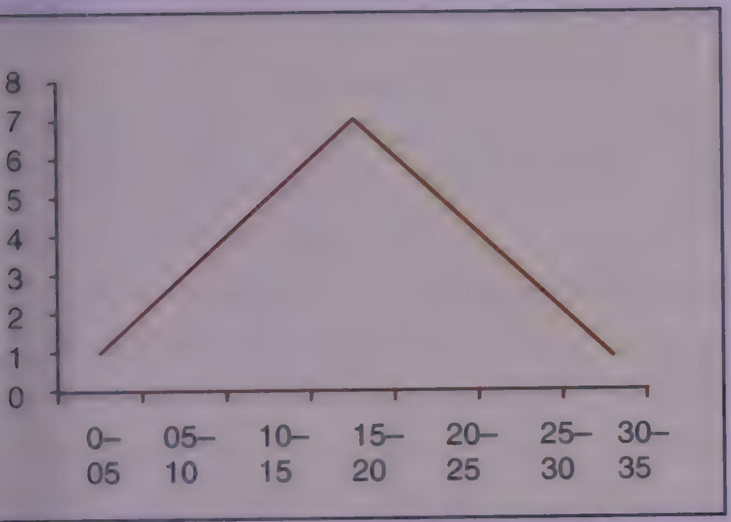
The standard deviation (along with several related measures like variance, coefficient of variati etc.) is used mostly in research studies and is regarded as a very satisfactory measure of dispers in a series. It is amenable to mathematical manipulation because the algebraic signs are not igno in its calculation (as we ignore in case of mean deviation). It is less affected by fluctuations sampling. These advantages make standard deviation and its coefficient a very popular measure the scatteredness of a series. It is popularly used in the context of estimation and testing of hypothes

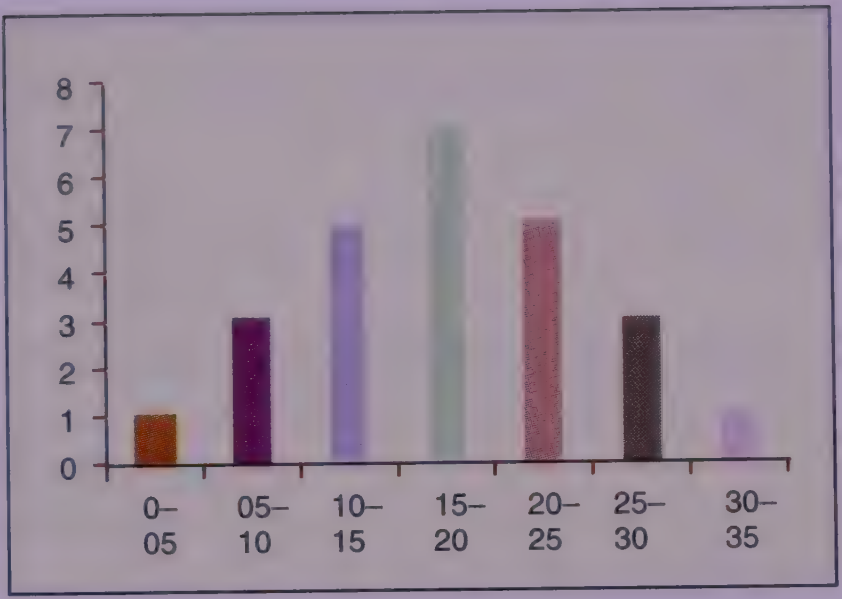
8.3 MEASURES OF SKEWNESS

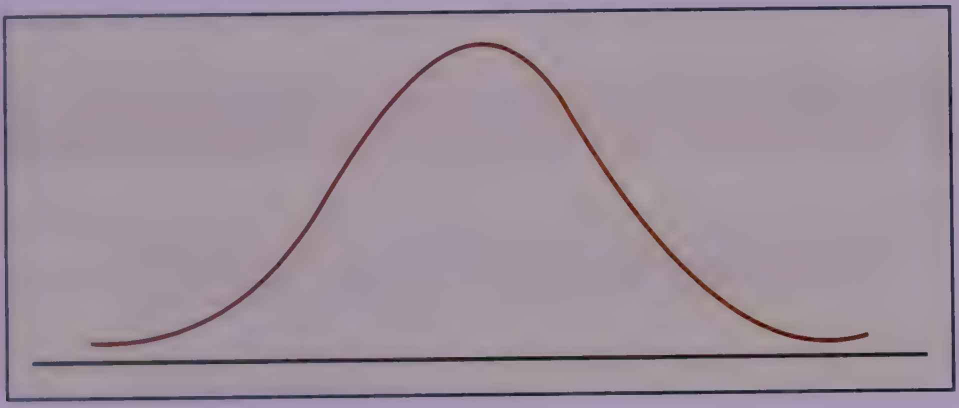
The meaning of skewness is lack of symmetry. Skewness gives us the idea of the shape of the distribut of the data. We can observe the shape of the distribution of the data-set by plotting a line chart histogram. Followings are the examples of symmetric data-sets when skewness is not present:



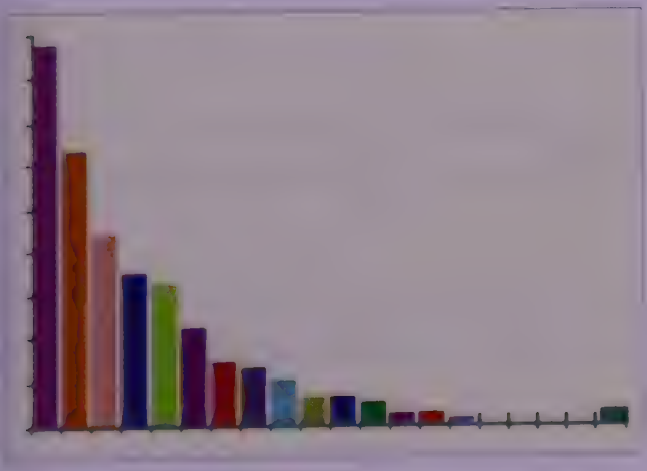


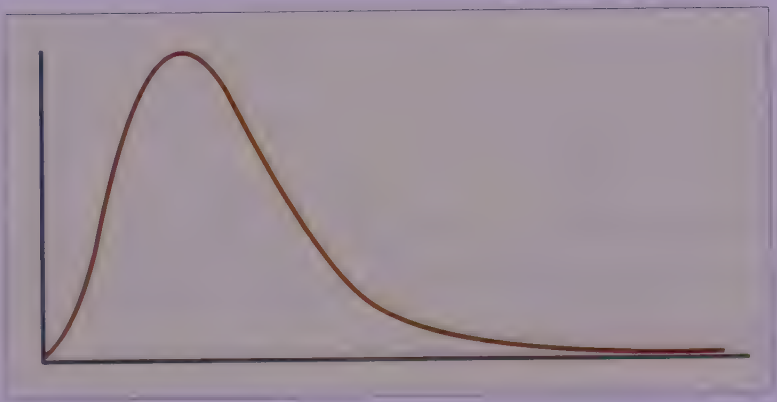






A data-set has a skewed distribution when mean, median, and mode (if exists) are not the same. uch a case, the plot of the distribution is stretched to one side than to the other. When the curve tretched towards right side more, we have positive skewness, and when the curve is stretched vards the left side more, we have negative skewness.





The difference between the mean, median or the mode provides an easy way of expressing wness of a data-set. In case of positive skewness, we have Mode < Median < Mean, and in case negative skewness, we have Mean < Median < Mode. Skewness is measured by (Mean – Mode). ase mode is ill-defined, it can be estimated from mean and median, for a moderately asymmetrical tribution, using the formula Mode = 3 Median – 2 Mean. Thus, when mode is ill-defined, skewness can be measured by 3(Mean – Median). In order to compare the skewness of two data-sets, obtain relative measures of skewness, as in dispersion. These measures are called as coefficient skewness and are pure numbers independent of units of measurements. A formula of coefficien skewness is

$$S_k = \frac{(\text{Mean–Mode})}{\text{Standard Deviation}}$$

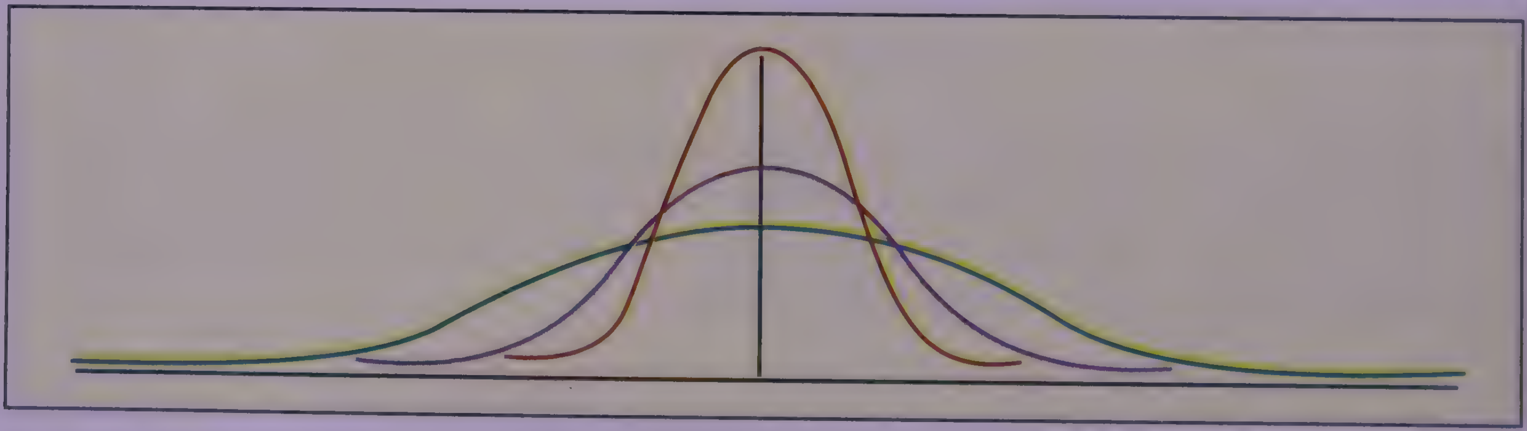
In case mode is ill-defined, the coefficient of skewness is

$$S_k = \frac{3(\text{Mean–Median})}{\text{Standard Deviation}}$$

The limits of this coefficient of skewness are ±3. For positive skewness, we have $0 < S_k < 3$, a for negative skewness, we have $-3 < S_k < 0$.

8.4 KURTOSIS

In addition to the central tendency, dispersion, and skewness, one more characteristic of the data– is measured sometimes which is called as kurtosis. It tells us about the flatness or the peakedness the distribution curve.



In the above figure, we have three distributions having smallest or flattest peak (platykutti having medium peak (mesokurtic), and having highest peak (leptokurtic). Measuring kurtosis is beyo the scope of this book.

It may be pointed out here that knowing the shape of the distribution curve is crucial to the use statistical methods in research analysis since most methods make specific assumptions about t nature of the distribution curve.

8.5 MEASURES OF RELATIONSHIP

So far we have dealt with those statistical measures that we use in context of univariate populati i.e., the population consisting of measurement of only one variable. But if we have the data on tv variables, we are said to have a bivariate population and if the data happen to be on more than tv iables, the population is known as multivariate population. Suppose we collect the data for monthly ome and expenditure of the individuals in a group. For each individuals in the group, we have a ue for income (variable X) and expenditure (variable Y). Thus, we have a pair of values for each ividual. This is an example of bivariate population. Similarly, we can have a multivariate population o. Suppose, we have the data for experience (variable Z) of the individuals in the same group. w we have a triplet of values for each individual in the group. Corresponding population is a ltivariate population.

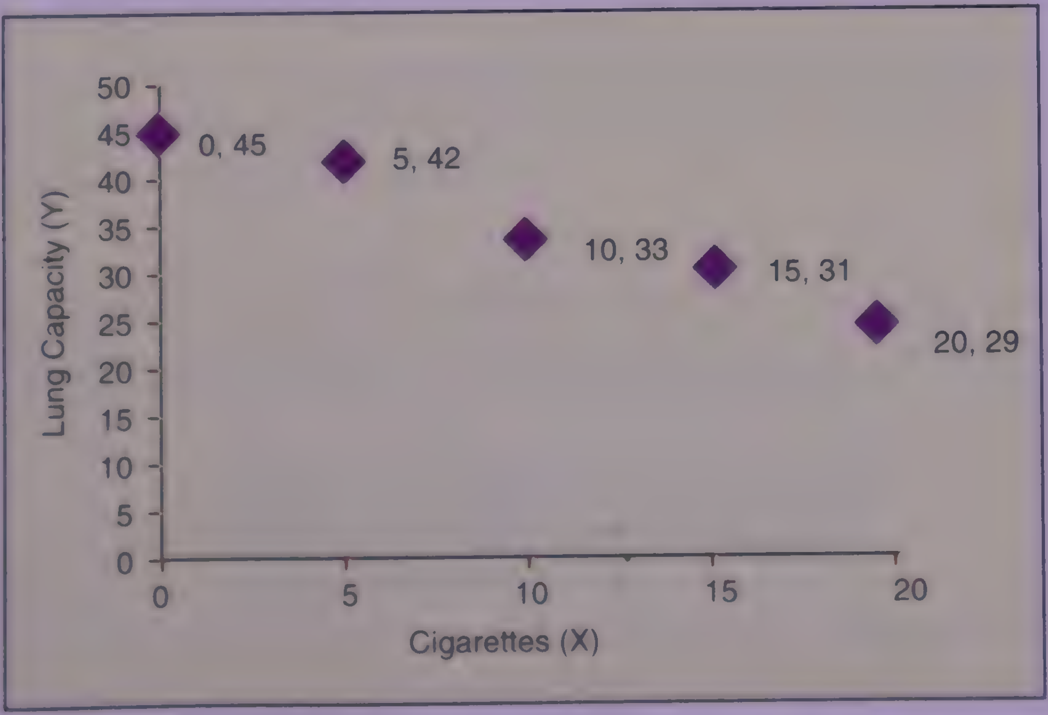
When the population is based on two or more characteristics (variables), we may like to measure relationship between the variables. The measurement of relationship between two or more variables give us some idea of the effect of one variable on the other.

A scatter plot is a useful graphical representation to have some approximate idea of relationship ween two variables. Scatter plot is obtained by plotting the pairs of observations taking one iable on X-axis and other on Y-axis. Suppose, for example, we want to investigate the relationship ween cigarette smoking and lung capacity. We might ask a group of people about their smoking oits, and measure their lung capacities. Following table represents the collected observations:

Table 8.1

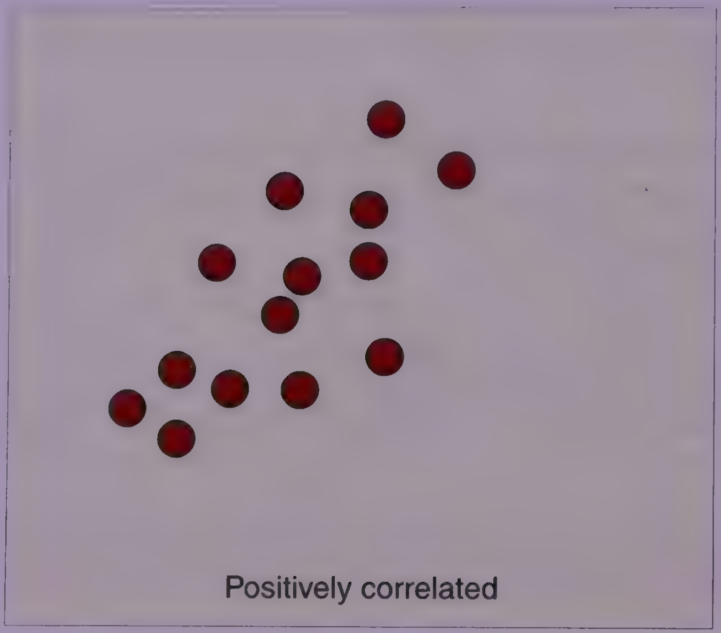
Cigarettes (X)	*Lung Capacity (Y)*
0	45
5	42
10	33
15	31
20	29

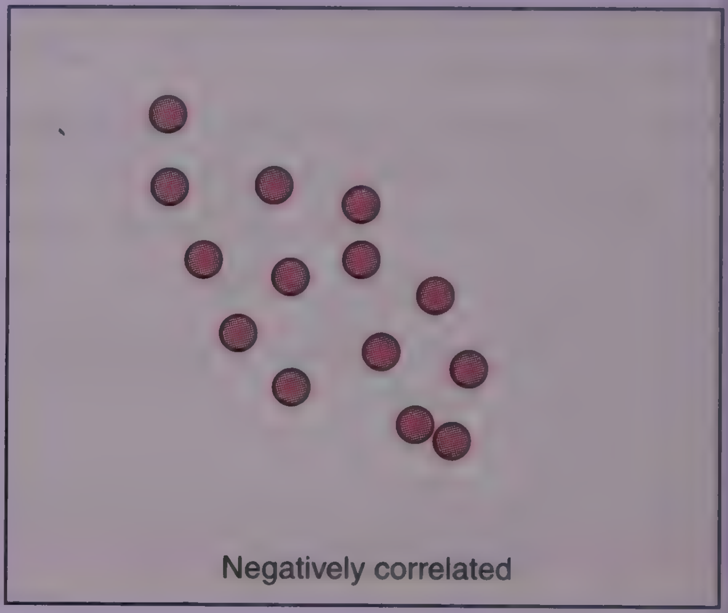
The scatter plot representation of this data is presented below:

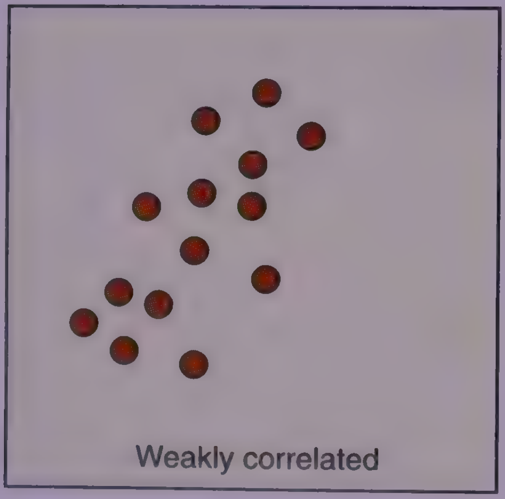


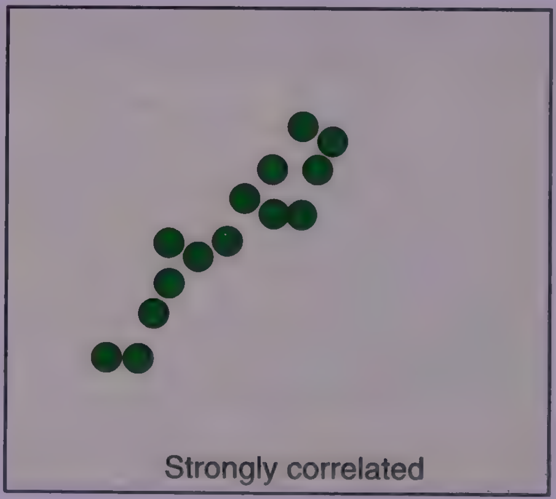
We can see that as smoking goes up, lung capacity tends to go down. The two variables change values in the opposite directions.

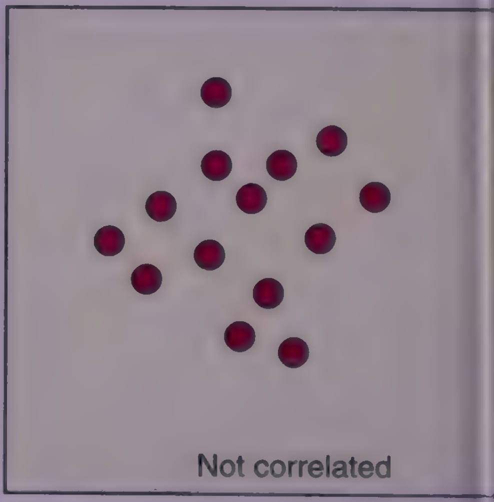
When the two variables change values in the opposite direction, we have negative correla When they change values in the same direction, they are said to be positively correlated. S scatter plots are presented below exhibiting various types of relationships:


Positively correlated


Negatively correlated


Weakly correlated


Strongly correlated


Not correlated

Some important measures of relationship are given below:

8.5.1 Covariance

In case of a bivariate population, where two variables are represented by X and Y and the pa observations are represented by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, then the covariance between X and

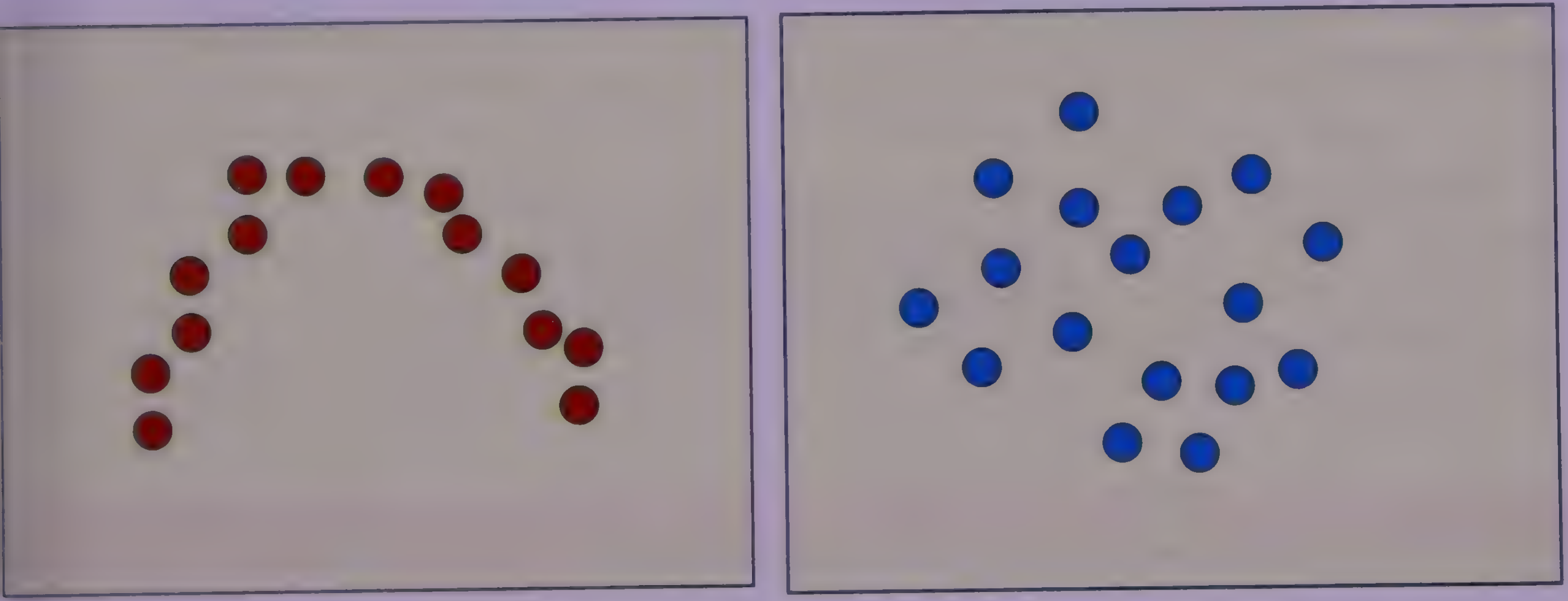
$$Cov(X,Y) = \sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

operties of Covariance

(i) $Cov(X+a,\ Y+b) = Cov(X,\ Y)$, where a and b are any constants. This means when some constant is added to all X values and the same or different constant is added to all Y values, covariance does not change. It is not affected by change in location.

(ii) $Cov(aX,\ bY) = ab\ Cov(X,\ Y)$. This means when all X values are multiplied by some constant and all Y values are multiplied by the same or different constant, covariance is changed. It is affected by change in scale.

(iii) Covariance can take any value from $-\infty$ to $+\infty$. Negative values indicate negative relationship, while positive values indicate positive relationship between the variables. Since covariance measures only linear relationship, so zero covariance indicates no linear relationship. In case of zero covariance, there may be some non-linear relationship or no relationship at all as shown in figure given below.



Also note that when all the observations on either variable of X and Y are the same, the covariance tween X and Y will be zero.

When we change the unit of observations on X or/and Y, the strength of relationship between and Y does not change, however the value of covariance is changed (property ii). So, covariance nnot be considered a good measure of relationship between two variables.

A better measure is correlation coefficient.

5.2 Karl Pearson's Coefficient of Correlation

arl Pearson's coefficient of correlation between the variables X and Y is given by

$$\boxed{r = \frac{Cov(X,Y)}{STDEV(X).STDEV(Y)} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}}$$

Karl Pearson's coefficient of correlation is not affected by change in scale or by change in cation. Unlike covariance it can be used to compare the relationships between two pairs of variables. is a unit free measure of relationship between two variables and takes values in $[-1, +1]$. When r

is close to +1 (−1), there is strong positive (negative) relationship. Like covariance, it also measu only linear relationship. It is also clear from the formula that correlation coefficient is zero whene covariance is zero.

8.5.3 Rank Correlation

Rank correlation or Charles Spearman's coefficient is the techniques of determining the degree correlation between two variables in case of ordinal data where ranks are given to the differ values of the variables. The main objective of this coefficient is to determine the extent to which two sets of ranking are similar or dissimilar. This coefficient is determined as under:

$$\text{Spearman's coefficient of correlation (or } r_s\text{)} = 1 - \left[\frac{6\sum d_i^2}{n\left(n^2 - 1\right)}\right]$$

where, d_i = Difference between ranks of *i*th pair of the two variables

n = Number of pairs of observations.

This measure is useful when a group of n individuals is arranged in order of merit or proficien in possession of two characteristics. Observations on each characteristic are assigned ranks fro 1 to n. If some individuals have the same merit or proficiency of a characteristic, they are assign the same rank (tied rank) which is the arithmetic mean of their positions after arranging them in risi order of the merit.

8.6 ASSOCIATION IN CASE OF ATTRIBUTES

When data is collected on the basis of some attribute or attributes, we have statistics commo termed as statistics of attributes. It is not necessary that the objects may process only one attribu rather it would be found that the objects possess more than one attribute. In such a situation o interest may remain in knowing whether the attributes are associated with each other or not. F example, among a group of people we may find that some of them are inoculated against small-p and among the inoculated we may observe that some of them suffered from small-pox after inoculati The important question which may arise for the observation is regarding the efficiency of inoculati for its popularity will depend upon the immunity which it provides against small-pox. In other wor we may be interested in knowing whether inoculation and immunity from small-pox are associate Technically, we say that the two attributes are associated if they appear together in a greater numb of cases than is to be expected if they are independent and not simply on the basis that they a appearing together in a number of cases as is done in ordinary life.

The association may be positive or negative (negative association is also known as disassociatio If class frequency of *AB*, symbolically written as (*AB*), is greater than the expectation of *AB* bei together if they are independent, then we say the two attributes are positively associated; but if t class frequency of *AB* is less than this expectation, the two attributes are said to be negative associated. In case the class frequency of *AB* is equal to expectation, the two attributes are considere as independent i.e., are said to have no association. It can be put symbolically as shown hereunde

If $(AB) > \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then AB are positively related/associated.

If $(AB) < \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then AB are negatively related/associated.

If $(AB) = \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then AB are independent i.e., have no association.

ere, (AB) = Frequency of class AB and

$\frac{(A)}{N} \times \frac{(B)}{N} \times N$ = Expectation of AB, if A and B are independent, and N being the number of items.

In order to find out the degree or intensity of association between two or more sets of attributes, should work out the coefficient of association. Professor Yule's coefficient of association is most pular and is often used for the purpose. It can be mentioned as under:

$$\boxed{Q_{AB} = \frac{(AB)(ab) - (Ab)(aB)}{(AB)(ab) + (Ab)(aB)}}$$

ere, Q_{AB} = Yule's coefficient of association between attributes A and B.

(AB) = Frequency of class AB in which A and B are present.

(Ab) = Frequency of class Ab in which A is present but B is absent.

(aB) = Frequency of class aB in which A is absent but B is present.

(ab) = Frequency of class ab in which both A and B are absent.

The value of this coefficient will be somewhere between +1 and –1. If the attributes are completely sociated (perfect positive association) with each other, the coefficient will be +1, and if they are mpletely disassociated (perfect negative association), the coefficient will be –1. If the attributes completely independent of each other, the coefficient of association will be 0. The varying degrees the coefficients of association are to be read and understood according to their positive and gative nature between +1 and –1.

Sometimes the association between two attributes, A and B, may be regarded as unwarranted hen we find that the observed association between A and B is due to the association of both A and with another attribute C. For example, we may observe positive association between inoculation d exemption for small-pox, but such association may be the result of the fact that there is positive sociation between inoculation and richer section of society and also that there is positive association tween exemption from small-pox and richer section of society. The sort of association between A d B in the population of C is described as *partial association* as distinguished from *total association* tween A and B in the overall universe. We can workout the coefficient of partial association tween A and B in the population of C by just modifying the above stated formula for finding sociation between A and B as shown below:

$$Q_{AB.C} = \frac{(ABC)(abC)-(AbC)(aBC)}{(ABC)(abC)+(AbC)(aBC)}$$

where,

$Q_{AB.C}$ = Coefficient of partial association between A and B in the population of C; and all o values are the class frequencies of the respective classes (A, B, C denotes the prese of concerning attributes and a, b, c denotes the absence of concerning attributes).

At times, we may come across cases of *illusory association*, wherein association between attributes does not correspond to any real relationship. This sort of association may be the resul some attribute, say C with which attributes A and B are associated (but in reality there is no associat between A and B). Such association may also be the result of the fact that the attributes A an might not have been properly defined or might not have been correctly recorded. Researcher m remain alert and must not conclude association between A and B when in fact there is no s association in reality.

In order to judge the significance of association between two attributes, we make use *Chi-square test** by finding the value of Chi-square (χ^2) and using Chi-square distribution the va of (χ^2) can be worked out as under:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad i = 1, 2, 3 \ldots \text{ and } j = 1, 2, 3$$

where, O_{ij} = Observed frequencies

E_{ij} = Expected frequencies.

Association between two attributes in case of manifold classification and the resulting continge table can be studied as explained below:

We can have manifold classification of the two attributes in which case each of the two attribu are first observed and then each one is classified into two or more subclasses, resulting into wha called as contingency table. The following is an example of 4 × 4 contingency table with two attribu A and B, each one of which has been further classified into four sub-categories.

Table 8.2: 4 x 4 Contingency Table

		Attribute A				
		A_1	A_2	A_3	A_4	*Total*
Attribute B	B_1	$(A_1 B_1)$	$(A_2 B_1)$	$(A_3 B_1)$	$(A_4 B_1)$	(B_1)
	B_2	$(A_1 B_2)$	$(A_2 B_2)$	$(A_3 B_2)$	$(A_4 B_2)$	(B_2)
	B_3	$(A_1 B_3)$	$(A_2 B_3)$	$(A_3 B_3)$	$(A_4 B_3)$	(B_3)
	B_4	$(A_1 B_4)$	$(A_2 B_4)$	$(A_3 B_4)$	$(A_4 B_4)$	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

---

* See Chapter "Chi-square test" for all details.

Association can be studied in a contingency table through Yule's coefficient of association as ted above, but for this purpose we have to reduce the contingency table into 2 × 2 table by mbining some classes. For instance, if we combine $(A_1) + (A_2)$ to form (A) and $(A_3) + (A_4)$ to form and similarly if we combine $(B_1) + (B_2)$ to form (B) and $(B_3) + (B_4)$ to form (b) in the above ntingency table, then we can write the table in the form of a 2 × 2 table as shown in Table 8.3.

Table 8.3

		Attribute		
		A	*a*	*Total*
Attribute	*B*	*(AB)*	*(aB)*	*(B)*
	b	*(Ab)*	*(ab)*	*(b)*
	Total	*(A)*	*(a)*	*N*

After reducing a contingency table in a two-by-two table through the process of combining some asses, we can work out the association as explained above. But the practice of combining classes not considered very correct and at times it is inconvenient also, Karl Pearson has suggested a easure known as *Coefficient of mean square contingency* for studying association in contingency bles. This can be obtained as under:

$$\boxed{C = \sqrt{\frac{\chi^2}{\chi^2 + N}}}$$

here, C = Coefficient of contingency

χ^2 = Chi-square value which is $= \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

N = Number of items.

This is considered a satisfactory measure of studying association in contingency tables.

.7 OTHER MEASURES

.7.1 Index Numbers

hen series are expressed in same units, we can use averages for the purpose of comparison, but hen the units in which two or more series are expressed happen to be different, statistical averages nnot be used to compare them. In such situations we have to rely upon some relative measurement hich consists in reducing the figures to a common base. Once such method is to convert the series to a series of index numbers. This is done when we express the given figures as percentages of some specific figure on a certain data. We can, thus, define an index number as a number whic used to measure the level of a given phenomenon as compared to the level of the same phenome at some standard date. The use of index number weights more as a special type of average, me to study the changes in the effect of such factors which are incapable of being measured direc But one must always remember that index numbers measure only the relative changes.

Changes in various economic and social phenomena can be measured and compared thro index numbers. Different indices serve different purposes. Specific commodity indices are to se as a measure of changes in the phenomenon of that commodity only. Index numbers may meas cost of living of different classes of people. In economic sphere, index numbers are often termec 'economic barometers measuring the economic phenomenon in all its aspects either directly measuring the same phenomenon or indirectly by measuring something else which reflects upon main phenomenon.

But index numbers have their own limitations with which researcher must always keep hims aware. For instance, index numbers are only approximate indicators and as such give only a fair ic of changes but cannot give an accurate idea. Chances of error also remain at one point or the ot while constructing an index number but this does not diminish the utility of index numbers for they s can indicate the trend of the phenomenon being measured. However, to avoid fallacious conclusio index numbers prepared for one purpose should not be used for other purposes or for the sa purpose at other places.

8.7.2 Time Series

In the context of economic and business researches, we may obtain quite often data relating to so time period concerning a given phenomenon. Such data is labelled as 'Time Series'. More clearly can be stated that series of successive observations of the given phenomenon over a period of ti are referred to as time series. Such series are usually the result of the effects of one or more of t following factors:

(i) *Secular trend* or long term trend that shows the direction of the series in a long period time. The effect of trend (whether it happens to be a growth factor or a decline factor) gradual, but extends more or less consistently throughout the entire period of time unc consideration. Sometimes, secular trend is simply stated as trend (or *T*).

(ii) *Short time oscillations* i.e., changes taking place in the short period of time only and su changes can be the effect of the following factors:

  (a) *Cyclical fluctuations (or C)* are the fluctuations as a result of business cycles and a generally referred to as long term movements that represent consistently recurri rises and declines in an activity.

  (b) *Seasonal fluctuations (or S)* are of short duration occurring in a regular sequence specific intervals of time. Such fluctuations are the result of changing seasons. Usual these fluctuations involve patterns of change within a year that tend to be repeat from year to year. Cyclical fluctuations and seasonal fluctuations taken together constitute short-period regular fluctuations.

(c) *Irregular fluctuations (or I)*, also known as Random fluctuations, are variations which take place in a completely unpredictable fashion.

All these factors stated above are termed as components of time series and when we try to alyse time series, we try to isolate and measure the effects of various types of these factors on a ies. To study the effect of one type of factor, the other type of factor is eliminated from the series. e given series is, thus, left with the effects of one type of factor only.

For analysing time series, we usually have two models; (1) multiplicative model; and (2) additive odel. Multiplicative model assumes that the various components interact in a multiplicative manner produce the given values of the overall time series and can be stated as under:

$$\boxed{Y = T \times C \times S \times I}$$

ere, Y = Observed values of time series
T = Trend
C = Cyclical fluctuations
S = Seasonal fluctuations
I = Irregular fluctuations.

Additive model considers the total of various components resulting in the given values of the erall time series and can be stated as:

$$\boxed{Y = T + C + S + I}$$

There are various methods of isolating trend from the given series viz., the free hand method, mi-average method, method of moving averages, method of least squares and similarly there are ethods of measuring cyclical and seasonal variations and whatever variations are left over are nsidered as random or irregular fluctuations.

The analysis of time series is done to understand the dynamic conditions for achieving the short-rm and long-term goals of business firm(s). The past trends can be used to evaluate the success or ilure of management policy or policies practiced hitherto. On the basis of past trends, the future tterns can be predicted and policy or policies may accordingly be formulated. We can as well study operly the effects of factors causing changes in the short period of time only, once we have iminated the effects of trend. By studying cyclical variations, we can keep in view the impact of clical changes while formulating various policies to make them as realistic as possible. The owledge of seasonal variations will be of great help to us in taking decisions regarding inventory, oduction, purchases and sales policies so as to optimize working results. Thus, analysis of time ries is important in context of long term as well as short term forecasting and is considered a very owerful tool in the hands of business analysts and researchers.

PROBLEMS

1. When is median preferred over mean to represent a set of values? Give example.
2. Explain the usage of mode by giving example.
3. Explain the requirement of a measure of dispersion.
4. Which one is the most preferred measure of dispersion and why?
5. Write short notes on skewness and kurtosis of a data-set.
6. Write short notes on geometric mean and harmonic mean of a data-set.
7. Why correlation coefficient is better than covariance as a measure of relationship between two variables
8. In the data set given in Advertisements.xls, obtain mean, median, mode and standard deviation of Sale (Column A), No. of Advertisements (Column B), and Expenditure on advertisements (Column C).
9. In the data-set given in Advertisements.xls, obtain covariance and correlation coefficients between No of Advertisements (Column B) and Expenditure on advertisements (Column C). What do you conclude
10. In the data-set given in BananaChips.xls, ranking of 3 bands of banana chips are given on price, qualit availability, packaging and taste by 15 respondents. Obtain average ranking on taste and quality an corresponding standard deviations.
11. In the data-set given in BananaChips.xls, obtain covariance and correlation coefficients between tast and quality. What do you conclude?

9

Chapter

Sampling and Statistical Inference

ampling is defined as the selection of some part of an aggregate or totality on the basis of which a udgement or inference about the aggregate or totality is made. In other words, it is the process of btaining information about an entire population by examining only a part of it. In most of the research vork and surveys, the usual approach happens to be to make generalisations or to draw inferences ased on samples about the parameters of population from which the samples are taken. The esearcher quite often selects only a few items from the universe for his study purposes. All this is lone on the assumption that the sample data will enable him to estimate the population parameters. The items so selected constitute what is technically called a sample, their selection process or technique s called sample design and the survey conducted on the basis of sample is described as sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

Sampling is used in practice for a variety of reasons such as:

1. Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.
2. Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.
3. Sampling remains the only way when population contains infinitely many members.
4. Sampling remains the only choice when a test involves the destruction of the item under study.
5. Sampling usually enables to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population.

A detailed discussion on the need for sampling and various sampling techniques is presented in Chapter 4.

9.1 PARAMETER AND STATISTIC

The collection of all the items about which the information (on one or more characteristics) is desir is called as population. For example, if we want to measure the average cellphone bill of the peo in a particular city in a particular month, our population would consist of the specified month's cellph bills of all the people in that city. The population or universe can be *finite* or *infinite*. The populati is said to be finite if it consists of a fixed number of elements so that it is possible to enumerate it its totality. For instance, the population of a city, the number of workers in a factory are examples finite populations. The symbol '*N*' is generally used to indicate how many elements (or items) a there in case of a finite population. An infinite population is that population in which it is theoretica impossible to observe all the elements. Thus, in an infinite population the number of items is infin i.e., we cannot have any idea about the total number of items. The number of stars in a sky, possib rolls of a pair of dice are examples of infinite population. One should remember that no truly infini population of physical objects does actually exist in spite of the fact that many such populatio appear to be very very large. From a practical consideration, we then use the term infinite populatio for a population that cannot be enumerated in a reasonable period of time. This way we use th theoretical concept of infinite population as an approximation of a very large finite population.

Any characteristic or measure of population units is known as a parameter. Population mea population standard deviation and population proportion are commonly studied parameters. Thes parameters are denoted by μ, σ and π respectively. Parameters are usually unknown as we do n always study the whole population. Unknown parameter are estimated by studying a subpart of th whole population which is known as a sample.

Any characteristic or measure of sample items is known as a statistic. Sample mean $(\bar{X})$ sample standard deviation (s_1) and sample proportion (p) are the examples of statistics. Obtainin the estimate of an unknown parameter using a statistic and studying the properties of the obtaine estimate are the prime objective of statistical inference.

9.2 SAMPLING AND NON-SAMPLING ERRORS

Sample surveys do imply the study of a small portion of the population and as such there woul naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be termed as sampling error or error variance. In other words, sampling errors are those errors which arise on account of sampling and they generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values. The meaning of sampling error can be easily understood from the Fig. 9.1.

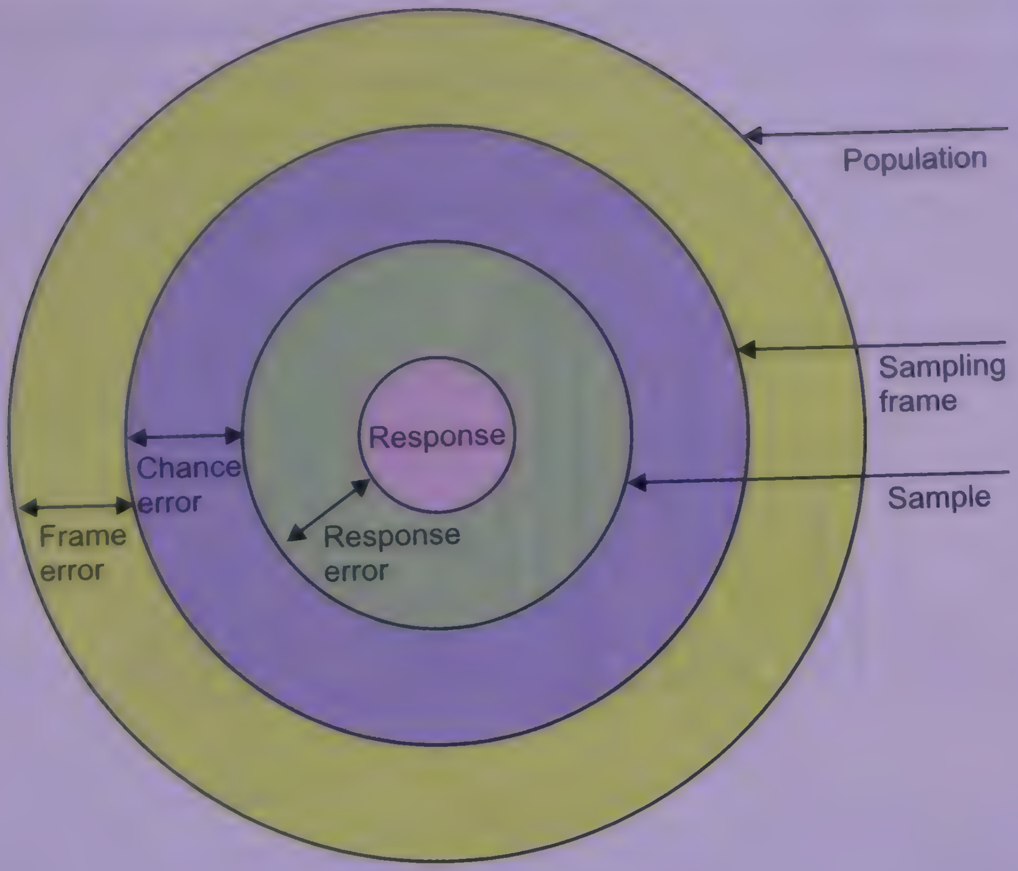


Fig. 9.1

Sampling error = Frame error + Chance error + Response error

(If we add measurement error or the non-sampling error to sampling error, we get total error).

Sampling errors occur randomly and are equally likely to be in either direction. The magnitude of he sampling error depends upon the nature of the universe; the more homogeneous the universe, maller the sampling error. Sampling error is inversely related to the size of the sample i.e., sampling rror decreases as the sample size increases and vice-versa. A measure of the random sampling rror can be calculated for a given sample design and size and this measure is often called the recision of the sampling plan. Sampling error is usually worked out as the product of the critical alue at a certain level of significance and the standard error.

As opposed to sampling errors, we may have non-sampling errors which may creep in during the rocess of collecting actual information and such errors occur in all surveys whether census or ample. We have no way to measure non-sampling errors.

9.3 SAMPLING DISTRIBUTION

Distribution of a statistic may not be the same as the distribution of population. We are often concerned with sampling distribution in sampling analysis. If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure.

Here, it is important to discuss normal distribution briefly. Normal distribution is represented the bell shape of the histogram of given data-set as below :

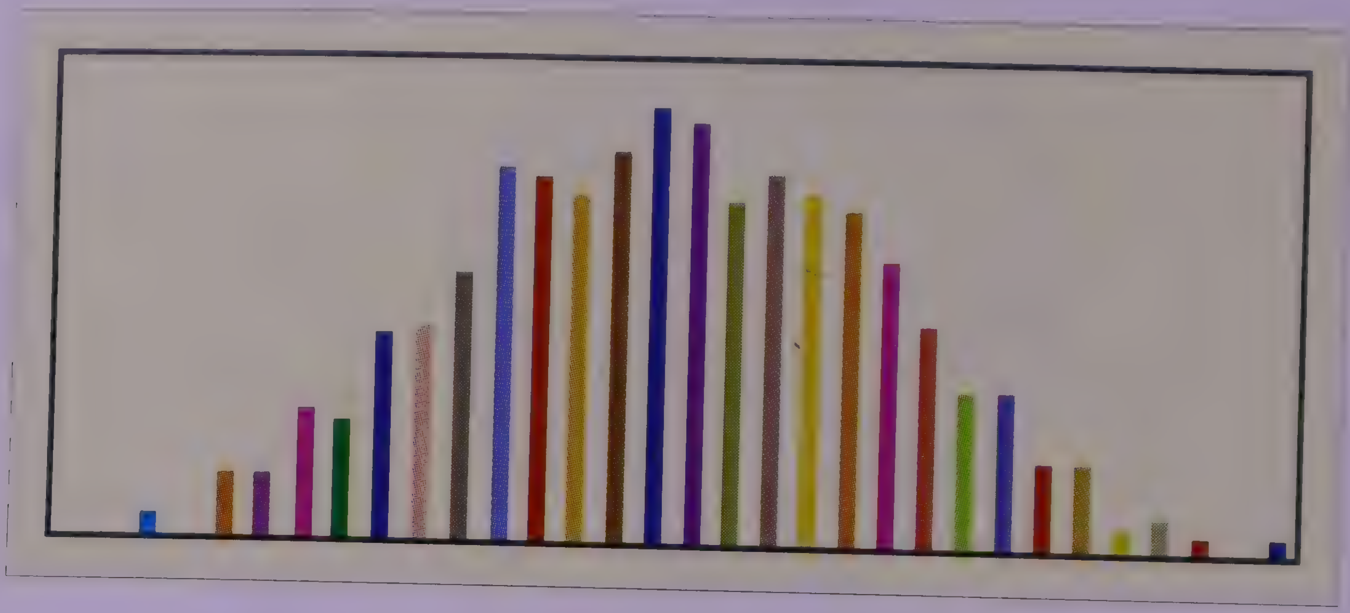


Fig. 9.2

In many statistical analyses, it is important that the given data has normal distribution. Norma distribution is denoted by $N(\mu,\sigma)$, where μ is mean and σ is standard deviation of the distributio Normal distribution can be used to model any variable taking values $-\infty$ to ∞, provided the values o that variable have a bell shaped histogram. Reader may refer to some standard text on statistics t know more about normal distribution.

Some commonly used sampling distributions are given below:

9.3.1 Sampling Distribution of Mean

Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population $N(\mu,\sigma)$, the sampling distribution of mean would also be normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation $=\sigma\sqrt{n}$, where μ is the mean of the population, σ is the standard deviation of the population and n means the number of items in a sample. But when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., $N(0, 1)$, we can write the normal variate $z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ for the sampling distribution of mean. This characteristic of the sampling distribution of mean is very useful in several decision situations.

9.3.2 Sampling Distribution of Proportion

Proportion is a measure of attribute. Let us consider that the population is divided into two mutually exclusive and collectively exhaustive classes—one class possessing a particular attribute while er class not possessing that attribute. For example, people in a city could be divided into "Smokers" l "Non-smokers". Let N = population size, X= number of people out of N possessing a particular ibute. $\pi = X/N$ = actual proportion of the people possessing the specified attribute. Let a sample elected from this population with n = sample size, x = number of people in the sample possessing specified particular attribute, $p = x/n$ = sample proportion. Note that, X and P are population ameters, while x and p are sample statistics. Also, p provides an estimate of π. It can be showed t the distribution of x is Binomial(n, π). Using the property of binomial distribution, for sufficiently ge sample size n, we have

$$Z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0,1)$$

Practically, this result is true for $n \geq 30$ or, when $n\pi \geq 5$ as well as $n(1-\pi) \geq 5$.

3.3 Student's *t*-Distribution

hen population standard deviation (σ) is not known and the sample is of a small size (i.e., $n \leq 30$), e use t-distribution for the sampling distribution of mean and workout t-variable as:

$= \frac{\bar{x} - \mu}{s_1/\sqrt{n}}$, where $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is sample variance. Student's t-distribution (or listribution) is also symmetrical and is very close to the distribution of standard normal variate, z, cept for small values of n. The variable t differs from z in the sense that we use sample standard eviation (s_1) in the calculation of t, whereas we use standard deviation of population (σ) in the lculation of z. There is a different t-distribution for every possible sample size i.e., for different egrees of freedom. The degrees of freedom for a sample of size n is $n-1$. As the sample size gets rger, the shape of the t-distribution becomes apporximately equal to the normal distribution. In fact or sample sizes of more than 30, the t-distribution is so close to the normal distribution that we can se the normal to approximate the t-distribution. Like normal distribution , the range of t-distribution also $(-\infty, \infty)$. The applications of t-distribution are discussed in further chapters.

.3.4 Chi-square (χ^2) Distribution

hi-square distribution is encountered when we deal with collections of values that involve adding up quares. Variances of samples require us to add a collection of squared quantities and thus have istributions that are related to chi-square distribution. Suppose we have a random sample $x_1, x_2, \ldots,$ $_n$ from $N(\mu, \sigma)$ population, then the distribution of the statistic $\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$ is chi-square distribution with n degree of freedom. Alternatively, the distribution of the statistic $\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{(n-1)s_1^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$ s chi-square distribution with $(n-1)$ degree of freedom. Clearly, the range of chi-square distribution includes all non-negative values. Chi square distribution is used in taking decisions about the varian testing independence of attributes and in testing goodness of fit. These applications of chi-squ distributions are discussed in further chapters.

9.3.5 Snedecor's *F*-Distribution

Let X and Y be two independent sample statistics such that X has chi-square distribution with degree of freedom and Y has chi-square distribution with d_2 degree of freedom. Then distribution the statistic $F = \dfrac{X/d_1}{Y/d_2}$ is Snedecor's F-distribution with d_1 and d_2 degree of freedom. The range this distribution is $(0, \infty)$.

Alternatively, suppose that we have two independent samples $(x_1, x_2, \ldots, x_{n_1})$ and $(y_1, y_2, \ldots,$ of sizes n_1 and n_2 respectively from normal populations having the same variance. Then distributi of the statistic $F = \dfrac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 / (n_1 - 1)}{\sum_{j=1}^{n_2} (y_j - \bar{y})^2 / (n_2 - 1)} = \dfrac{s_x^2}{s_y^2}$ is Snedecor's F-distribution with $(n_1 - 1)$ a $(n_2 - 1)$ degree of freedom. Here if s_x^2 and s_y^2 are the sample variances. F-ratio is computed in way that the larger variance is always in the numerator. F-distribution is used in comparing tw variances and analysis of variance technique discussed later.

9.4 DEGREE OF FREEDOM

The number of independent observations which make up a statistic is known as the degrees freedom (d.f.) associated with that statistic. Degree of freedom is the number of values in the fina calculation of a statistic that are free to vary. In general, d.f. of a statistic = Number of independen observations – Number of parameters estimated. For example, d.f. of the statistic $\sum_{i=1}^{n} \left(\dfrac{x_i - \bar{x}}{\sigma}\right.$ is $(n - 1)$, while d.f. of the statistic $\sum_{i=1}^{n} \left(\dfrac{x_i - \mu}{\sigma}\right)^2$ is n.

9.5 STANDARD ERROR

We have seen that different samples of the same size from the same population will yield differen values of statistic under consideration, say sample means. A measure of the variability in differen values of sample mean is given by the Standard Error of the sample mean. Standard error of a statistic is the standard deviation of its sampling distribution. Standard error plays an important role ir statistical hypothesis testing and interval estimation. Standard error gives an idea about the reliability and precision of the estimate. Smaller the standard error, greater the uniformity of sampling distribution and hence, greater the reliability of the estimate. Also, standard error decreases when sample size is increased. Standard errors of some important statistics are given below:

Table 9.1

Statistic	Standard Error
Sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\sigma/\sqrt{n}$ (σ is population standard deviation)
Sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$	$\sigma^2\sqrt{2/n}$
Sample variance $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$	$\sigma^2\sqrt{2/(n-1)}$
Sample proportion $p = x/n$	$\sqrt{\pi(1-\pi)/n}$

As can be noticed that standard error depends on (usually) unknown parameters. In order to get ome approximated value of standard error, unknown population parameters can be replaced by ome feasible estimates.

9.6 CENTRAL LIMIT THEOREM

hen sampling is from a normal population, the means of samples drawn from such a population are emselves normally distributed. But when sampling is not from a normal population, the size of the ample plays a critical role. When n is small, the shape of the distribution will depend largely on the hape of the parent population, but as n gets large ($n > 30$), the shape of the sampling distribution will ecome more and more like a normal distribution, irrespective of the shape of the parent population. he theorem which explains this sort of relationship between the shape of the population distribution nd the sampling distribution of the mean is known as the central limit theorem. This theorem is by ar the most important theorem in statistical inference. It assures that the sampling distribution of the ean approaches normal distribtion as the sample size increases. In formal terms, we may say that e central limit theorem states that the distribution of means of random samples taken from a opulation having mean μ and finite variance σ^2 approaches the normal distribution with mean μ and ariance σ^2/n as n goes to infinity.

"The significance of the central limit theorem lies in the fact that it permits us to use sample tatistics to make inferences about population parameters without knowing anything about the shape f the frequency distribution of that population.

For example, 1,800 randomly selected values from an exponential distribution are plotted below:

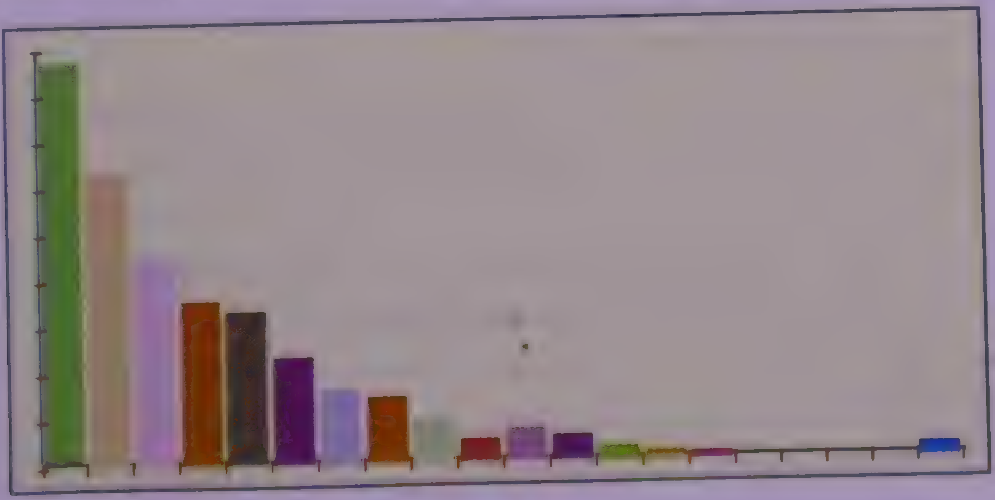


Fig. 9.3

Various samples of sizes 2, 5, and 30 are drawn from this population and sample mean values obtained. The distributions of the sample means for various sample sizes can be observed fr following plots:

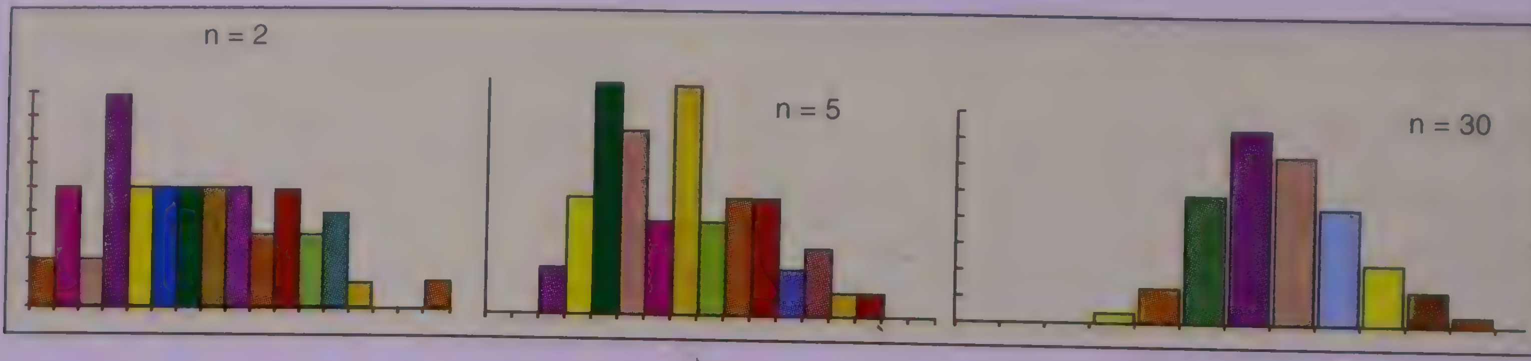


Fig. 9.4

It can be noted that for sample size 30 the shape of sampling distribution is almost normal.

9.7 FINITE POPULATION CORRECTION

For the application of central limit theorem, we assumed that the sample size n is large. If t population size N is small, sample size n cannot be sufficiently large and we cannot apply central lim theorem. In this situation, we multiply the standard error by finite population correction (*fpc*), whi is given by $fpc = \sqrt{\frac{N-n}{N-1}}$. Clearly, when $N \to \infty, fpc \to 1$. In practice, *fpc* may be used wh $n/N > 0.05$.

9.8 STATISTICAL INFERENCE

A statistical population is a collection of items or individuals about which we wish to draw som conclusion. As discussed, population may be finite or infinite. It may also be hypothetical or existen For example, consider the random experiment of tossing a coin till we get the first head. Her population is infinite and imaginary. Existent population is a population of concrete objects, e.g number of bulbs in a box. Whenever we do not study the whole population to make conclusions abou some characteristics of that population, we study a sample and using the sample we make conclusion about the unknown characteristics of that population. Statistics is helpful in studying the sample an drawing meaningful inference from the sample. This requires probabilistic sample as statistical result are based on probability theory. Drawing inference from the probabilistic sample about unknow population parameters is called as statistical inference. Statistical inference is concerned mainly wit two things—Hypothesis testing and Estimation.

In hypothesis testing we test the claims made about unknown population parameters using sample These claims are made using some past experiences or beliefs. This topic is discussed in detail i further chapters.

Estimation means estimating unknown population parameters using a sample. There are tw types of estimates—point estimate and interval estimate.

9.8.1 Point Estimation

Point estimate is a single valued estimate of an unknown parameter. The statisti $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean) is used to estimate population mean μ. So, sample mean is a point imate of the population mean. The statistics $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ and $s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ used to estimate unknown population variance σ^2. Therefore, these two statistics s^2 and s_i^2 are nt estimates of σ^2. Conventionally, $\bar{x}$, s^2 and s_1^2 are called as estimators and specific values ch as $\bar{x}=20, s^2=4$) are called as estimates of the parameters. An unknown parameter may be imated by more than one estimator. For example, σ^2 is estimated by s^2 and s_1^2. Out of these two imators which one is better? To answer this question we must know about the properties of a good imator as given below:

Unbiasedness

t us consider the following population of size four: 18, 20, 22, and 24. Clearly,
pulation mean = (18 + 20 + 22 + 24)/ 4 = 21 and
pulation variance = $[(18-21)^2+(20-21)^2+(22-21)^2+(24-21)^2]/4=5$.
ow we collect all possible samples of size two from this population and calculate sample mean $=\frac{1}{n}\sum_{i=1}^{n}x_i$, sample variance $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ and modified sample variance $=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ for each sample. Here n is the sample size and x_i denotes the sample servations. These values are presented in the following table:

Table 9.2

Samples	$\bar{x}$	s^2	s_1^2
18, 18	18	0	0
20, 18	19	1	2
22, 18	20	4	8
24, 18	21	9	18
18, 20	19	1	2
20, 20	20	0	0
22, 20	21	1	2
24, 20	22	4	8
18, 22	20	4	8
20, 22	21	1	2
22, 22	22	0	0
24, 22	23	1	2
18, 24	21	9	18
20, 24	22	4	8
22, 24	23	1	2
24, 24	24	0	0
Average	21	2.5	5

We can notice that average of all sample mean values is the same as population mean 21. average of all s^2 values is not the same as population variance, while the average of all s_1^2 value the same as population variance 5.

This property of an estimate is known as unbiasedness. An estimate, say t, of a parameter, sa is known as an unbiased estimate of θ if the average of all possible values of t (obtained from possible samples of a given size) is the same as θ. This is denoted as $E(t) = \theta$. Equivalently, expected value of t (long run or overall average of t values) should be the same as the value o From the example we can observe that sample mean is an unbiased estimate of population me Sample variance s^2 is not an unbiased estimate of population variance, while Sample variance s_1 an unbiased estimate of population variance. This is written as E (sample mean) = population me $E(s^2) \neq$ population variance, and $E(s_1^2)$ = population variance.

(ii) *Consistency*

An estimator should approach the value of population parameter as the sample size becomes lar and larger. This property is referred to as the *property of consistency*. This is the most desira property of an estimator. When sample size becomes the same as the population size, sample becom the population. In such a case sample mean becomes population mean. Also Sample variance becomes population variance σ^2. For sufficiently large sample size n the values of s^2 and s_1^2 a almost the same. Thus sample mean is a consistent estimator of population mean. Both samp variances s^2 and s_1^2 are consistent estimators of population variance σ^2.

(iii) *Sufficiency*

An estimator should use as much as possible the information available from the sample. This proper is known as the *property of sufficiency*. The details on this property are beyond the scope of th book.

(iv) *Efficiency*

An estimator should have a relatively small variance. This means that the most efficient estimato among a group of unbiased estimators, is one which has the smallest variance. This property technically described as the *property of efficiency*.

Keeping in view the above stated properties, the researcher must select appropriate estimator(s) f his study. Below we present the point estimate of some important parameters.

9.8.2 Interval Estimation

In interval estimation, we obtain using a sample statistic at a desired level of confidence, e.g., 99 confidence. Meaning of 99% confidence is the probability that the obtain interval will possess th true unknown parameter is 0.99. This confidence cannot be 100% as the confidence interval in th case will be infinite $((-\infty, \infty) \text{ or } (0, \infty))$. In point estimation we give a single value while in interva mation we give an interval to estimate a parameter. In interval estimation sample to sample ation is also taken into consideration through standard error. General rule of obtaining interval mate is described below:

Let θ be the unknown parameter and let T be the unbiased point estimate of θ, i.e., $E(T) = \theta$. w fix a desired level of confidence denoted as $(1 - \alpha) \times 100\%$. Here α is the probability of pe I error (to be discussed later). Usually confidence level is fixed as 90%, 95% or 99%. The ue $(1 - \alpha)$ is called confidence coefficient. Thus, for 95% confidence level, $\alpha = 0.05$. The confidence erval estimate of θ is $[T - h, T + h]$. It means that

$$\boxed{P(T - h \leq \theta \leq T + h) = 1 - \alpha,}$$

ere, h = Critical value x standard error. Critical Value is obtained using confidence coefficient $-\alpha)$ and the sampling distribution of point estimator T. In the interval estimate, $T - h$ is called as ver confidence limit and $T + h$ is called as upper confidence limit.

Estimating Population Mean μ

en σ^2 is known:

As discussed, in order to obtain interval estimate of a parameter, we require

(a) the unbiased point estimate of that parameter,
(b) standard error of the point estimate,
(c) confidence coefficient (to be fixed by the researcher),
(d) critical value which will be obtained by sampling distribution of the point estimate and confidence coefficient.

The unbiased point estimate of population mean μ is sample mean $\bar{x}$. The standard error of $\bar{x}$ is $\sqrt{n}$. We have already discussed that the distribution of $\dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ is N(0, 1) provided the parent pulation is normal or sample size n is sufficiently large. Under this condition, critical value $z_{\alpha/2}$ is ven by the following probability equation:

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = (1 - \alpha)$$

here Z has N(0, 1) distribution.

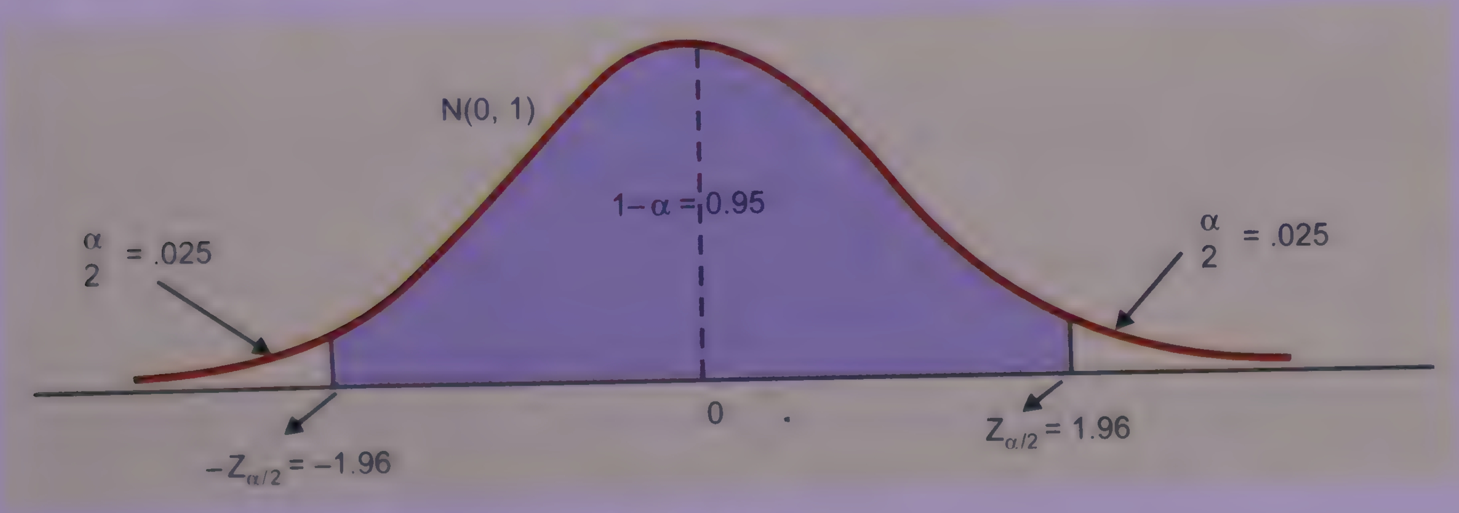


Fig. 9.5

Critical values for some commonly used confidence levels (for N(0, 1) distribution) are provi below:

Table 9.3

Confidence level	Confidence coefficient	α	Critical valu
80%	0.8	0.2	1.28
90%	0.9	0.1	1.645
95%	0.95	0.05	1.96
98%	0.98	0.02	2.33
99%	0.99	0.01	2.58
99.80%	0.998	0.002	3.08
99.90%	0.999	0.001	3.27

Thus $(1-\alpha)\times 100\%$ confidence interval estimate of μ is given by $\left(\bar{x} - z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}\right.$

Example 9.1: A sample of 11 circuits from a large normal population has a mean resistance of 2. ohms. We know from past testing that the population standard deviation is 0.35 ohms. Determin 95% confidence interval for the true mean resistance of the population.

Solution: 95% confidence interval for the true mean resistance of the population is given

$$\left(\bar{x} - z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}\right) = \left(2.20-1.96\times\frac{0.35}{\sqrt{11}},\ 2.20+1.96\times\frac{0.35}{\sqrt{11}}\right) = (2.20 - 0.2068,\ 2.$$

$+ 0.2068) = (1.9932, 2.4068)$. The meaning of this is we are 95% sure that the true population me is within (1.9932, 2.4068).

When σ^2 is not known:

In case when population variance σ^2 is not known it can be replaced by its unbiase estimate $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. The distribution of $\frac{\bar{x}-\mu}{s_1/\sqrt{n}}$ would be N(0, 1) when sample size large enough (in practice, $n \geq 30$), whatever be the distribution of parent population, normal or nor normal. Thus in case of large samples, (1 – α) × 100% confidence interval estimate of μ is given b

$$\left(\bar{x} - z_{\alpha/2}\times\frac{s_1}{\sqrt{n}},\ \bar{x}+z_{\alpha/2}\times\frac{s_1}{\sqrt{n}}\right).$$

However, in case of small samples, the distribution of $\frac{\bar{x}-\mu}{s_1/\sqrt{n}}$ is not normal. Because of som added variability in the statistic (as we are using s_1 in place of σ), the distribution of $\frac{\bar{x}-\mu}{s_1/\sqrt{n}}$ would b ent's t-distribution with $(n-1)$ d.f. (which is more widened than normal's bell shape) provided parent population is normal. In such situation, $(1-\alpha)\times 100\%$ confidence interval estimate of μ

ven by $\left(\overline{x}-t_{\alpha/2}\times\frac{s_1}{\sqrt{n}},\ \overline{x}+t_{\alpha/2}\times\frac{s_1}{\sqrt{n}}\right)$ where, the critical value $t_{\alpha/2}$ is given by t distribution's

e for known confidence level and d.f. such that $P(-t_{\alpha/2}<T<t_{\alpha/2})=(1-\alpha)$, for $T\sim t_{(n-1)}$. Some ical Values of $t_{(n-1)}$ distribution for given α and d.f. $(n-1)$ are given below:

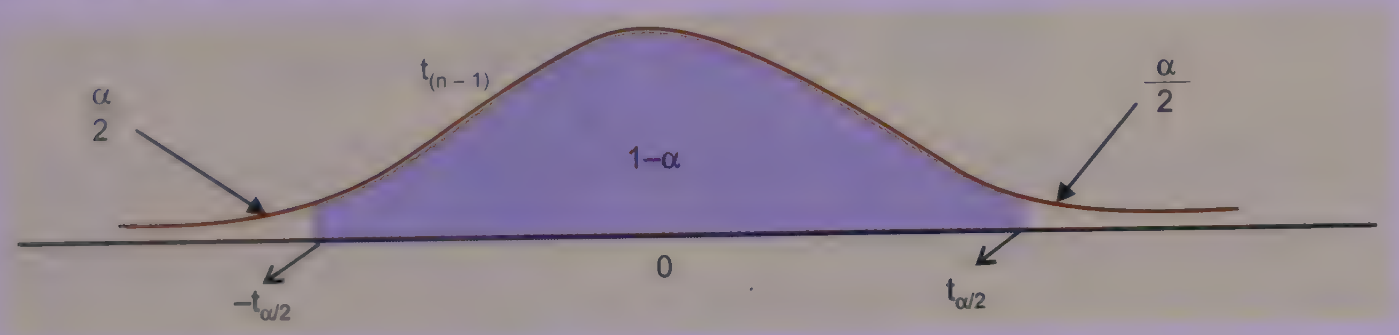


Fig. 9.6

Table 9.4

d.f. $(n-1)$	Critical value at $\alpha = 0.05$	Critical value at $\alpha = 0.10$
1	12.706	6.314
2	4.303	2.92
3	3.182	2.353
4	2.776	2.132
5	2.571	2.015
6	2.447	1.943
7	2.365	1.895

For more critical values related to t-distribution, we refer Table 2 in the Appendix.

ample 9.2: A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ns. Population standard deviation is not known. Sample standard deviation (s_1) is 0.35 ohms. ermine a 95% confidence interval for the true mean resistance of the population.

ution: 95% confidence interval for the true mean resistance of the population is given by

$$-t_{\alpha/2}\times\frac{s_1}{\sqrt{n}},\ \overline{x}+t_{\alpha/2}\times\frac{s_1}{\sqrt{n}}\Bigg)=\left(2.20-2.228\times\frac{0.35}{\sqrt{11}},\ 2.20+2.228\times\frac{0.35}{\sqrt{11}}\right)$$

$$=(2.20-0.235,\ 2.20+0.235)=(1.965,\ 2.435).$$

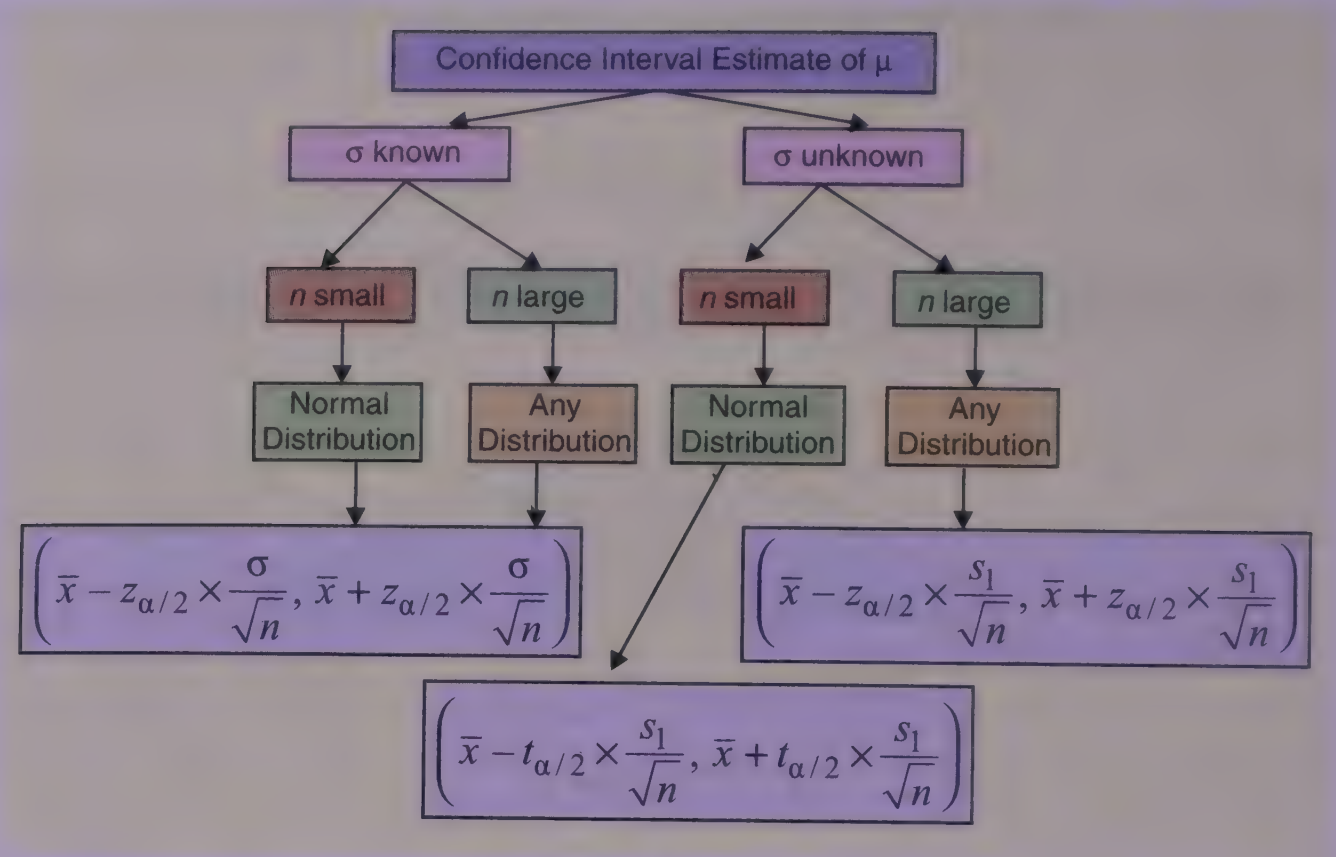

Confidence Interval Estimate of μ
σ known
σ unknown
n small
n large
n small
n large
Normal Distribution
Any Distribution
Normal Distribution
Any Distribution
$\left(\bar{x} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$
$\left(\bar{x} - z_{\alpha/2} \times \frac{s_1}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \times \frac{s_1}{\sqrt{n}}\right)$
$\left(\bar{x} - t_{\alpha/2} \times \frac{s_1}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \times \frac{s_1}{\sqrt{n}}\right)$

Fig. 9.7

(ii) Estimating Population Proportion π

As discussed an unbiased estimate of population proportion π is provided by sample proport $p{=}x/n$ and the standard error of p is $\sqrt{\pi(1-\pi)/n}$. We have also seen that the distributio

statistic $\dfrac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$ is N(0, 1) for sufficiently large sample size n. Using this, $(1-\alpha)\times 100\%$ confide

interval estimate of π is given by $\left(p - z_{\alpha/2} \times \sqrt{\pi(1-\pi)/n},\ p + z_{\alpha/2} \times \sqrt{\pi(1-\pi)/n}\right)$. But the probl with this expression is it itself contains π, which is unknown. Therefore, this confidence inter estimate becomes meaningless. In order to get a meaningful confidence interval estimate of π replace π by its unbiased estimate p. Thus, $(1-\alpha)\times 100\%$ confidence interval estimate of π $\left(p - z_{\alpha/2} \times \sqrt{p(1-p)/n},\ p + z_{\alpha/2} \times \sqrt{p(1-p)/n}\right)$, provided that the sample size is large enough practice, $n \geq 30$ or $n\pi \geq 5$ as well as $n(1-\pi) \geq 5$).

Example 9.3: A random sample of 100 people shows that 25 have opened IRA (individual retirem arrangement) this year. Construct a 95% confidence interval for the true proportion of people w have opened IRA.

Solution: 95% confidence interval for the true proportion of people who have opened IRA is giv

by $\left(p - z_{\alpha/2} \times \sqrt{\dfrac{p(1-p)}{n}}, p + z_{\alpha/2} \times \sqrt{\dfrac{p(1-p)}{n}}\right)$

$$= \left(0.25 - 1.96 \times \sqrt{0.25 \times \frac{0.75}{100}}, 0.25 + 1.96 \times \sqrt{0.25 \times \frac{0.75}{100}}\right)$$

$=(0.25 - 1.96 \times 0.0433, 0.25 + 1.96 \times 0.0433) = (0.25 - 0.08487, 0.25 + 0.0848$
$=(0.16513, 0.33487)$

Estimating Population Variance σ^2

ance is an inverse measure of a group's homogeneity. It is an important indicator of total quality andardized products and services. A process is improved by reducing the variance. In financial agement, variance is a measure of financial risk. Variance of rates of return helps the managers ss financial and capital investment alternatives. Productivity, wages, and costs of living vary veen regions and nations. Because of wide usefulness of variance, it is important to estimate nown population variance. We have two point commonly used point estimates of population ance σ^2 given as $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$, $s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. We have already seen in Section l that s^2 is not an unbiased estimate while s_1^2 is unbiased for estimating σ^2. However both are sistent estimators of σ^2. For sufficiently large sample size, $s_1^2=s^2$.

Confidence interval estimate of population variance is obtained in two situations, viz. small sample large sample, separately as below:

se 1: Small sample (n < 30)

m Section 9.3.4, we know that the distribution of the statistic $\frac{(n-1)s_1^2}{\sigma^2}$ is chi-square distribution h $(n-1)$ degree of freedom provided the parent population is normal. Using this distribution, the - $\alpha)\times 100\%$ confidence interval estimate of σ^2 is given by $\left(\frac{(n-1)s_1^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s_1^2}{\chi^2_{1-\alpha/2}}\right)$, where

and $\chi^2_{1-\alpha/2}$ are the critical values obtained using chi-square distribution with $(n-1)$ d.f. and fidence coefficient α. These critical values can be taken from Table 3 in Appendix. The critical ue $\chi^2_{\alpha/2}$ is the point on X-axis such that the area bounded by chi-square probability curve, $=\chi^2_{\alpha/2}$ and X = ∞ is $\alpha/2$. Similarly, the critical value $\chi^2_{1-\alpha/2}$ is the point on X-axis such that the a bounded by chi-square probability curve, X = $\chi^2_{1-\alpha/2}$ and X = ∞ is $(1-\alpha/2)$. Below we show the ical values for chi-square distribution with 7 d.f. at 90% confidence level on chi-square probability ve.

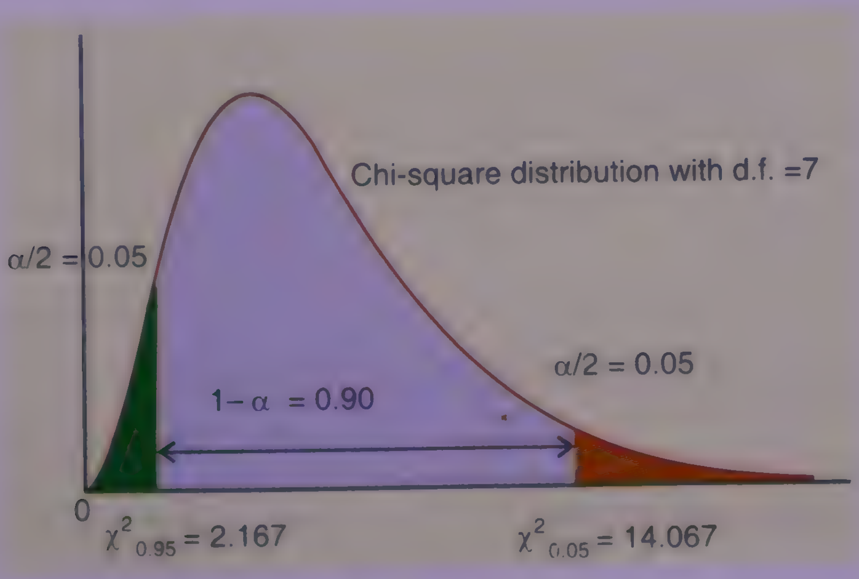


Fig. 9.8

Example 9.4: The cholesterol concentration in the yolks of a sample of 18 randomly selected
laid by genetically engineered chickens were found to have a mean value of 9.38 mg/g of yolk
standard deviation of 1.62 mg/g. Use this information to construct a 95% confidence interval esti
of the true variance of the cholesterol concentration in these egg yolks.

Solution: 90% confidence interval estimate of the true variance of the cholesterol concentrati
given by

$$\left(\frac{(n-1)s_1^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s_1^2}{\chi^2_{1-\alpha/2}}\right) = \left(\frac{17\times1.62\times1.62}{27.587}, \frac{17\times1.62\times1.62}{8.672}\right) = (1.61724, 5.144696).$$

Case 2: Large sample ($n \geq 30$)

Sample variance s_1^2 is unbiased for estimating population variance σ^2. Also the standard error o
is $\sigma^2\sqrt{2/(n-1)}$. By central limit theorem, for sufficiently large sample, the distributio

$\dfrac{s_1^2-\sigma^2}{\sigma^2\sqrt{2/(n-1)}}$ is N(0, 1). Using this we can obtain the $(1-\alpha)\times 100$ % confidence interval esti

of σ^2 as $\left(\dfrac{s_1^2}{1+z_{\alpha/2}\sqrt{2/(n-1)}}, \dfrac{s_1^2}{1-z_{\alpha/2}\sqrt{2/(n-1)}}\right)$. Here $z_{\alpha/2}$ is the critical value obtained u
N(0, 1) distribution.

Example 9.5: A technologist is developing a new method for processing a food material. For
quality, it is important to control moisture content in the final product. So, as one part of determi
the practicality of the new method, the technologist must estimate the variability of water conte
the resulting product. He collects 50 specimens of product from the new process, and determines
percent water in each. These 50 specimens give sample mean water content of 43.24% a
sample standard deviation of 7.93%. Compute a 95% confidence interval estimate of the true vari
of the percentage water for this new process.

Solution: 95% confidence interval estimate of the true variance is

$$\left(\frac{s_1^2}{1+z_{\alpha/2}\sqrt{2/(n-1)}}, \frac{s_1^2}{1-z_{\alpha/2}\sqrt{2/(n-1)}}\right) = \left(\frac{7.93\times7.93}{1+1.96\times\sqrt{2/49}}, \frac{7.93\times7.93}{1-1.96\times\sqrt{2/49}}\right)$$

$$=(45.04714, 104.1106).$$

9.8.3 Sample Size and its Determination

In sampling analysis the most ticklish question is: What should be the size of the sample or how la
or small should be '*n*'? If the sample size ('*n*') is too small, it may not serve to achieve the objecti
and if it is too large, we may incur huge cost and waste resources. As a general rule, one can say t
the sample must be of an optimum size i.e., it should neither be excessively large nor too sm

nically, the sample size should be large enough to give a confidence interval of desired width and ch the size of the sample must be chosen by some logical process before sample is taken from niverse. Size of the sample should be determined by a researcher keeping in view the following ts:

(i) *Nature of universe:* Universe may be either homogenous or heterogeneous in nature. If the items of the universe are homogeneous, a small sample can serve the purpose. But if the items are heterogeneous, a large sample would be required. Technically, this can be termed as the dispersion factor.

(ii) *Number of classes proposed:* If many class-groups (groups and sub-groups) are to be formed, a large sample would be required because a small sample might not be able to give a reasonable number of items in each class-group.

(iii) *Nature of study:* If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is considered appropriate in technical surveys.

(iv) *Type of sampling:* Sampling technique plays an important role in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

(v) *Standard of accuracy and acceptable confidence level:* If the standard of accuracy or the level of precision is to be kept high, we shall require relatively larger sample. For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.

(vi) *Availability of finance:* In practice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample for large samples result in increasing the cost of sampling estimates.

(vii) *Other considerations:* Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study are a few other considerations to which a researcher must pay attention while selecting the size of the sample.

There are two alternative approaches for determining the size of the sample. The first approach to specify the precision of estimation desired and then to determine the sample size necessary to sure it and the second approach uses the cost of additional information against the expected value the additional information. The first approach is capable of giving a mathematical solution, and as ch is a frequently used technique of determining '*n*'. The limitation of this technique is that it does t analyse the cost of gathering information *vis-a-vis* the expected value of information. The second proach is theoretically optimal, but it is seldom used because of the difficulty involved in measuring e value of information. Hence, we shall mainly concentrate here on the first approach.

8.3.1 Determination of Sample Size (Confidence Level Approach)

begin with, it can be stated that whenever a sample study is made, there arises some sampling ror which can be controlled by selecting a sample of adequate size. Researcher will have to ecify the precision that he wants in respect of his estimates concerning the population parameters.

For instance, a researcher may like to estimate the mean of the universe within ±3 of the true m
with 95 per cent confidence. In this case we will say that the desired precision is ±3, i.e., if t
sample mean is ₹ 100, the true value of the mean will be no less than ₹ 97 and no more than ₹
In other words, all this means that the acceptable error, *e*, is equal to 3. Keeping this in view, we
now explain the determination of sample size so that specified precision is ensured.

(a) *Sample size decision while estimating mean:* The confidence interval for the universe mean is given by

$$\boxed{\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

where, $\bar{X}$ = Sample mean;

$z_{\alpha\ 2}$ = The value of the standard variate at a given confidence level (to be read from table giving the areas under normal curve as shown in appendix) and it is 1.96 a 95% confidence level;

n = Size of the sample;

σ = Standard deviation of the popultion (to be estimated from past experience or on basis of a trial sample). Suppose, we have $\sigma = 4.8$ for our purpose.

If the difference between μ and $\bar{X}$ or the acceptable error is to be kept with in ±3 of the samp mean with 95% confidence, then we can express the acceptable error, '*e*' as equal to

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad 3 = 1.96\frac{4.8}{\sqrt{n}}$$

Hence, $n = \dfrac{(1.96)^2(4.8)^2}{(3)^2} = 9.834 \cong 10$

In a general way, if we want to estimate μ in a population with standard deviation σ with error no greater than '*e*' by calculating a confidence interval with confidence corresponding to z the necessary sample size, *n*, equals as under:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{e^2}$$

All this is applicable when the population happens to be infinite. But in case of finite populatio the above stated formula for determining sample size will be modified as

$$n^* = \frac{z^2_{\alpha/2} N\sigma^2}{(N-1)e^2 + z^2_{\alpha/2}\sigma^2}$$

ere,

N = Size of population

n = Size of sample

e = Acceptable error or the precision or the estimation error

σ = Standard deviation of the population

$z_{\alpha/2}$ = Standard varuate or z value or critical value using N(0, 1) distribution for confidence level α.

Proof (may be omitted from the discussion):

ample 9.6 : Determine the size of the sample for estimating the true weight of the cereal containers the universe with $N = 5000$ on the basis of the following information:

1. The variance of weight = 4 ounces on the basis of past records.
2. Estimate should be within 0.8 ounces of the true average weight with 99% probability.

---

*In case of finite population the confidence interval for μ is given by

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$

ere $\sqrt{(N-n)/(N-1)}$ is the finite population multiplier and all other terms mean the same thing as stated above.

If the precision is taken as equal to 'e' then we have

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$

or

$$e^2 = z^2_{\alpha/2}\frac{\sigma^2}{n} \times \frac{N-n}{N-1}$$

or

$$e^2(N-1) = \frac{z^2_{\alpha/2}\sigma^2 N}{n} - \frac{z^2_{\alpha/2}\sigma^2 n}{n}$$

or

$$e^2(N-1) + z^2_{\alpha/2}\sigma^2 = \frac{z^2_{\alpha/2}\sigma^2 N}{n}$$

or

$$n = \frac{z^2_{\alpha/2}\sigma^2 N}{e^2(N-1) + z^2_{\alpha/2}\sigma^2},$$

or

$$n = \frac{z^2_{\alpha/2} N\sigma^2}{e^2(N-1) + z^2_{\alpha/2}\sigma^2}$$

Will there be a change in the size of the sample if we assume infinite population in the g case? If so, explain by how much?

Solution: In the given problem we have the following:

$N = 5000$

$\sigma = 2$ ounces (since the variance of weight = 4 ounces);

$e = 0.8$ ounces (since the estimate should be within 0.8 ounces of the true average weig

$z_{\alpha/2} = 2.57$ (as per the table of area under normal curve for the given confidence level of 99

Hence, the confidence interval for μ is given by

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

and accordingly the sample size can be worked out as under:

$$n = \frac{z_{\alpha/2}^2 N \sigma^2}{(N-1)e^2 + z_{\alpha/2}^2 \sigma^2}$$

$$= \frac{(2.57)^2 \cdot (5000) \cdot (2)^2}{(5000-1)\,(0.8)^2 + (2.57)^2\,(2)^2}$$

$$= \frac{132098}{3199.36 + 26.4196} = \frac{132098}{3225.7796} = 40.95 \cong 41$$

Hence, the sample size (or n) = 41 for the given precision and confidence level in the abo question with finite population. But if we take population to be infinite, the sample size will be work out as under:

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{e^2}$$

$$= \frac{(2.57)^2\,(2)^2}{(0.8)^2} = \frac{26.4196}{0.64} = 41.28 \simeq 41$$

Thus, in the given case the sample size remains the same even if we assume infinite populatio

In the above illustration, the standard deviation of the population was given, but in many cas the standard deviation of the population is not available. Since we have not yet taken the sample a are in the stage of deciding how large to make it (sample), we cannot estimate the populaion standa deviation. In such a situation, if we have an idea about the range (i.e., the difference between t highest and lowest values) of the population, we can use that to get a crude estimate of the standa deviation of the population for getting a working idea of the required sample size. We can get the sa estimate of standard deviation as follows:

Since 99.7 per cent of the area under normal curve lies within the range of ±3 standard deviation we may say that these limits include almost all of the distribution. Accordingly, we can say that t given range equals 6 standard deviations because of ±3. Thus, a rough estimate of the populatio standard deviation would be:

$$6\hat{\sigma} = \text{The given range}$$

$$\hat{\sigma} = \frac{\text{The given range}}{6}$$

If the range happens to be, say ₹ 12, then

$$\hat{\sigma} = \frac{12}{6} = ₹\,2.$$

d this estimate of standard deviation, $\hat{\sigma}$, can be used to determine the sample size in the formulae ted above.

Sample size decision while estimating a percentage or proportion: If we are to find the mple size for estimating a proportion, our reasoning remains similar to what we have said in the ntext of estimating the mean. First of all, we shall have to specify the precision and the confidence el and then we will work out the sample size as under:

Since the confidence interval for universe proportion, $\hat{p}$ is given by

$$p \pm z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}}$$

ere, $p =$ Sample proportion, $q = 1 - p$;

$z_{\alpha/2} =$ The value of the standard variate at a given confidence level and to be worked out from table showing area under Normal Curve;

$n =$ Size of sample.

Since $\hat{p}$ is actually what we are trying to estimate, then what value we should assign to it? One ethod may be to take the value of $p = 0.5$ in which case 'n' will be the maximum and the sample ll yield at least the desired precision. This will be the most conservative sample size. The other ethod may be to take an initial estimate of p which may either be based on personal judgement or ay be the result of a pilot study. In this context it has been suggested that a pilot study of something e 225 or more items may result in a reasonable approximation of p value.

Then with the given precision rate, the acceptable error, 'e', can be expressed as under:

$$e = z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

$$e^2 = z_{\alpha/2}^2 \frac{pq}{n}$$

$$n = \frac{z_{\alpha/2}^2 \cdot p \cdot q}{e^2}.$$

The formula gives the size of the sample in case of infinite population when we are to estimate e proportion in the universe. But in case of finite population the above stated formula will be anged as under:

$$n = \frac{z_{\alpha/2}^{2} \cdot p \cdot q \cdot N}{e^{2}(N-1) + z_{\alpha/2}^{2} \cdot p \cdot q}$$

Example 9.7: What should be the size of the sample if a simple random sample from a populatio 4000 items is to be drawn to estimate the per cent defective within 2 per cent of the true value v 95.5 per cent probability? What would be the size of the sample if the population is assumed to infinite in the given case?

Solution: In the given question we have the following:

$N = 4000$;

$e = 0.02$ (since the estimate should be within 2% of true value);

$z_{\alpha 2} = 2.005$ (as per table of area under normal curve for the given confidence level of 95.5

As we have not been given the p value being the proportion of defectives in the universe, le assume it to be $p = 0.02$ (This may be on the basis of our experience or on the basis of past data may be the result of a pilot study).

Now we can determine the size of the sample using all this information for the given questio follows:

$$n = \frac{z_{\alpha/2}^{2} \cdot p \cdot q \cdot N}{e^{2}(N-1) + z_{\alpha/2}^{2} \cdot p \cdot q}$$

$$= \frac{(2.005)^2\ (.02)\ (1-.02)\ (4000)}{(.02)^2\ (4000-1) + (2.005)^2\ (.02)\ (1-.02)}$$

$$= \frac{315.1699}{1.5996 + .0788} = \frac{315.1699}{1.6784} = 187.78 \simeq 188$$

But if the population happens to be infinite, then our sample size will be as under:

$$n = \frac{z_{\alpha/2}^{2} \cdot p \cdot q}{e^{2}}$$

$$= \frac{(2.005)^2 \cdot (.02)\ (1-.02)}{(.02)^2}$$

$$= \frac{0.0788}{0.0004} = 196.98 \simeq 197.$$

Example 9.8: Suppose a certain hotel management is interested in determining the percentage the hotel's guests who stay for more than 3 days. The reservation manager wants to be 95 per ce confident that the percentage has been estimated to be within ±3% of the true value. What is th most conservative sample size needed for this problem?

tion: We have been given the following:

Population is infinite;

$e = 0.03$ (since the estimate should be within 3% of the true value);

$z_{\alpha/2} = 1.96$ (as per table of area under normal curve for the given confidence level of 95%).

ve want the most conservative sample size we shall take the value of $p = .5$ and $q = .5$. Using his information, we can determine the sample size for the given problem as under:

$$n = \frac{z_{\alpha/2}^2\, p\, q}{e^2}$$

$$= \frac{(1.96)^2 \cdot (.5)\,(1-.5)}{(.03)^2} = \frac{.9604}{.0009} = 1067.11 \simeq 1067$$

Thus, the most conservative sample size needed for the problem is = 1067.

Sample size decision while estimating variance: We have discussed that the distribution of

$$\frac{-\sigma^2}{\sqrt{\frac{2}{n-1}}}$$

is N (0, 1) for sufficiently large sample. If precision or estimation error $e = \left|s_1^2 - \sigma^2\right|$ is ed by the researcher in advance, then for fixed confidence coefficient $(1-\alpha)$.

$$z_{\alpha/2} = \frac{e}{\sigma^2\sqrt{2/(n-1)}}$$

Solving this we get,

$$n = 1 + \frac{2\sigma^4 z_{\alpha/2}^2}{e^2}$$

But above formula depends on unknown population variance σ^2. For the purpose of obtaining mple size here, a rough estimate of σ^2 can be used from some other small sample or pilot survey. no information is available, it can be decided by the researcher using past experience or can be en as the square of Range/6.

8.3.2 Determination of Sample Size (Bayesian Approach)

is approach of determining '*n*' utilises Bayesian statistics and as such is known as Bayesian approach. e procedure for finding the optimal value of '*n*' or the size of sample under this approach is as under:

(i) Find the expected value of the sample information (EVSI) for every possible *n*;

(ii) Also workout reasonably approximated cost of taking a sample of every possible *n*;

(iii) Compare the EVSI and the cost of the sample for every possible *n*. In other words, workout the expected net gain (ENG) for every possible *n* as stated below:

For a given sample size (*n*):

(EVSI) – (Cost of sample) = (ENG)

(iv) From (iii) above the optimal sample size, that value of n which maximises the differe between the EVSI and the cost of the sample, can be determined.

The computation of EVSI for every possible n and then comparing the same with the respec cost is often a very cumbersome task and is generally feasible with mechanised or computer h Hence, this approach although being theoretically optimal is rarely used in practice.

9.8.4 Tests of Significance (Hypothesis Testing)

An important aspect of sampling theory is the study of the tests of significance, which enable u decide on the basis of sample if the deviation between`

(i) the observed sample statistic and hypothetical parametric value, or

(ii) two independent sample statistics is significant or might be attributed to the fluctuation sampling the topic is discussed in further chapters in detail.

More Solved Examples

Example 9.9: From a random sample of 36 New Delhi civil service personnel, the mean age and sample standard deviation were found to be 40 years and 4.5 years respectively. Construct a 95 cent confidence interval for the mean age of civil servants in New Delhi.

Solution: The given information can be written as under:

$$n = 36$$

$$\overline{X} = 40 \text{ years}$$

$$s_1 = 4.5 \text{ years}$$

and the standard variate, z, for 95 per cent confidence is 1.96 (as per the normal curve area tabl

Thus, 95 per cent confidence inteval for the mean age of population is:

$$\overline{X} \pm z \frac{s_1}{\sqrt{n}}$$

or $$40 \pm 1.96 \frac{4.5}{\sqrt{36}}$$

or $$40 \pm (1.96)(0.75)$$

or $$40 \pm 1.47 \text{ years}$$

Example 9.10: In a random selection of 64 of the 2400 intersections in a small city, the me number of scooter accidents per year was 3.2 and the sample standard deviation was 0.8.

1. Make an estimate of the standard deviation of the population from the sample standa deviation.
2. Work out the standard error of mean for this finite population.

3 If the desired confidence level is 0.90, what will be the upper and lower limits of the confidence interval for the mean number of accidents per intersection per year?

ution: The given information can be written as under:

$N = 2400$ (This means that population is finite)

$n = 64$

$\overline{X} = 3.2$

$s_1 = 0.8$

the standard variate (z) for 90 per cent confidence is 1.645 (as per the normal curve area table).

Now we can answer the given questions thus:

1. The best point estimate of the standard deviation of the population is the standard deviation of the sample itself.

   Hence, $\hat{\sigma} = s_1 = 0.8$

2. Standard error of mean for the given finite population is as follows:

$$\sigma_{\overline{X}} = \frac{s_1}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{0.8}{\sqrt{64}} \times \sqrt{\frac{2400-64}{2400-1}}$$

$$= \frac{0.8}{\sqrt{64}} \times \sqrt{\frac{2336}{2399}}$$

$$= (0.1)\,(0.97)$$

$$= 0.097$$

3. 90 per cent confidence interval for the mean number of accidents per intersection per year is as follows

$$\overline{X} \pm z\left\{\frac{s_1}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right\}$$

$$= 3.2 \pm (1.645)\,(.097)$$

$= 3.2 \pm .16$ accidents per intersection.

ample 9.11: The foreman of *ABC* mining company has estimated the average quantity of iron e extracted to be 36.8 tons per shift and the sample standard deviation to be 2.8 tons per shift. sed upon a random selection of 4 shifts. Construct a 90 per cent confidence interval around this imate.

Solution: As the standard deviation of population is not known and the size of the sample is small shall use *t*-distribution for finding the required confidence interval about the population mean. given information can be written as under:

$$\overline{X} = 36.8 \text{ tons per shift}$$

$$s_1 = 2.8 \text{ tons per shift}$$

$$n = 4$$

degrees of freedom = $n - 1 = 4 - 1 = 3$ and the critical value of '*t*' for 90 per cent confidence inte or at 10 per cent level of significance is 2.353 for 3 *d.f.* (as per the table of *t*-distribution).

Thus, 90 per cent confidence interval for population mean is

$$\overline{X} \pm t\frac{s_1}{\sqrt{n}}$$

$$= 36.8 \pm 2.353\ \frac{2.8}{\sqrt{4}} = 36.8 \pm (2.353)\ (1.4)$$

$$= 36.8 \pm 3.294 \text{ tons per shift.}$$

Example 9.12: A market research survey in which 64 consumers were contacted states that 64 cent of all consumers of a certain product were motivated by the product's advertising. Find confidence limits for the proportion of consumers motivated by advertising in the population, giv confidence level equal to 0.95.

Solution: The given information can be written as under:

$n = 64$

$p = 64\%$ or 0.64

$q = 1 - p = 1 - 0.64 = 0.36$

and the standard variate (z) for 95 per cent confidence is 1.96 (as per the normal curve area tab

Thus, 95 per cent confidence interval for the proportion of consumers motivated by advertisin the population is:

$$p \pm z \cdot \sqrt{\frac{pq}{n}}$$

$$= .64 \pm 1.96\ \sqrt{\frac{(0.64)\ (0.36)}{64}}$$

$$= 0.64 \pm (1.96)(0.06)$$

$$= 0.64 \pm 0.1176$$

Thus, lower confidence limit is 52.24%.

Upper confidence limit is 75.76%.

mple 9.13: A firm has a population of 1000 accounts and wishes to estimate the total population e. A sample of 80 accounts is selected with average balance of $87.6 and standard deviation of 3. Find the 95% confidence interval estimate of the total balance.

ution: $N = 1000,\ n = 80, \bar{x} = 87.6, s_1 = 22.3$

confidence interval estimate of the total balance is

$$N\bar{x} \pm N\, z_{0.025} \frac{s_1}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$$= (1000)(87.6) \pm (1000)(1.96) \frac{22.3}{\sqrt{80}} \sqrt{\frac{1000-80}{1000-1}}$$

$$= 87{,}600 \pm 4{,}762.48$$

$$= (82837.52,\ 92362.48)$$

mple 9.14: Econe Dresses has 1200 inventory items. In the past 15% items were incorrectly ed. A sample of 120 items was selected. Historical cost of each item was compared with the ited value. 15 items differ in their historical costs and audited values. These values are as follows:

Table 9.5

Historical Cost	*Audited Value*	*D_i = (Historical Cost – Audited Value)*
261	240	21
87	105	–18
201	276	–75
121	110	11
315	298	17
411	356	55
249	211	38
216	305	–89
21	210	–189
140	152	–12
129	112	17
340	216	124
341	402	–61
135	97	38
228	220	8

Obtain 95% confidence interval estimate of total difference.

Solution: We first calculate the differences as given in the above table. Sample size $n = 120$. sample has 15 non-zero values given in the table above. Rest 105 values are zero in the sam Using the sample of 120 items, we calculate sample mean $\bar{D} = -0.95833$, and sample stan deviation as $s_D = 25.24482$. Population size is given as $N = 1200$. Since $n/N = 120/1200 = 0.1 >$ we use *fpc*. 95% confidence interval estimate of total difference is given by

$$\left(N\bar{D} \pm N z_{(0.025)} \frac{s_D}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right)$$

$$= \left[1200 \times (-0.95833) \pm 1200 \times 1.96 \times \frac{25.24482}{\sqrt{120}} \sqrt{\frac{1200-120}{1200-1}} \right]$$

Table 9.6 : Summarized Formulae (Interval Estimation)

Population Mean (μ)	σ is known	Small sample (Normal Distribution)	$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
		Large sample (Any Distribution)	
	σ is not known	Small sample (Normal Distribution)	$\bar{x} \mp t_{\alpha/2} \times \frac{s_1}{\sqrt{n}}$
		Large sample (Any Distribution)	$\bar{x} \mp z_{\alpha/2} \times \frac{s_1}{\sqrt{n}}$
Population Proportion (P)		Small sample	Out of scope
		Large sample (Any Distribution)	$p \mp z_{\alpha/2} \times \sqrt{pq/n}$
Population Variance $(\sigma)^2$		Small sample (Normal Distribution)	$\frac{(n-1)\, s_1^2}{\chi^2_{\alpha/2}}, \frac{(n-1)\, s_1^2}{\chi^2_{1-\alpha/2}}$
		Large sample (Any Distribution)	$\frac{s_1^2}{1 \pm z_{\alpha/2} \sqrt{2/(n-1)}}$

Table 9.7 : Summarized Formulae (Sample Size Determination)

For estimating Population Mean (μ)	Large sample (Any Distribution)	$n=\left(\frac{\sigma z_{\alpha/2}}{e}\right)^2$
For estimating Population Proportion (P)	Large sample (Any Distribution)	$n=\frac{PQ\ (z_{\alpha/2})^2}{e^2}$
For estimating Population Variance $(\sigma)^2$	Large sample (Any Distribution)	$n=1+\frac{2\sigma^4 z_{\alpha/2}^2}{e^2}$

PROBLEMS

1. Explain the meaning and significance of the concept of "Standard Error' in sampling analysis.
2. Describe briefly the commonly used sampling distributions.
3. State the reasons why sampling is used in the context of research studies.
4. Explain the meaning of the following sampling fundamentals:
   (a) Sampling frame;
   (b) Sampling error;
   (c) Central limit theorem;
   (d) Student's *t* distribution;
   (e) Finite population multiplier.
5. Distinguish between the following:
   (a) Statistic and parameter;
   (b) Confidence level and significance level;
   (c) Random sampling and non-random sampling;
   (d) Sampling of attributes and sampling of variables;
   (e) Point estimate and interval estimation.
6. Write a brief essay on statistical estimation.
7. 500 articles were selected at random out of a batch containing 10000 articles and 30 were found defective. How many defective articles would you reasonably expect to find in the whole batch?
8. In a sample of 400 people, 172 were males. Estimate the population proportion at 95% confidence level.
9. A sample of 16 measurements of the diameter of a sphere gave a mean $\overline{X} = 4.58$ inches and a standard deviation $\sigma_s = 0.08$ inches. Find (a) 95%, and (b) 99% confidence limits for the actual diameter.
10. A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the standard error of the population of bad ones in a sample of this size is 0.015 and also show that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

11. From a packet containing iron nails, 1000 iron nails were taken at random and out of them 100 were fou defective. Estimate the percentage of defective iron nails in the packet and assign limits within whi the percentage probably lies.

12. A random sample of 200 measurements from an infinite population gave a mean value of 50 an standard deviation of 9. Determine the 95% confidence interval for the mean value of the populatio

13. In a random sample of 64 mangoes taken from a large consignment, some were found to be bad. Dedu that the percentage of bad mangoes in the consignment almost certainly lies between 31.25 and 68. given that the standard error of the proportion of bad mangoes in the sample 1/16.

14. A random sample of 900 members is found to have a mean of 4.45 cms. Can it be reasonably regarded a sample from a large population whose mean is 5 cms and variance is 4 cms?

15. It is claimed that Americans are 16 pounds overweight on average. To test this claim, 9 random selected individuals were examined and the average excess weight was found to be 18 pounds. At th 5% level of significance, is there reason to believe the claim of 16 pounds to be in error?

16. The foreman of a certain mining company has estimated the average quantity of ore extracted to be 34 tons per shift and the sample standard deviation to be 2.8 tons per shift, based upon a random selectio of 6 shifts. Construct 95% as well as 98% confidence interval for the average quantity of ore extracte per shift.

17. A sample of 16 bottles has a mean of 122 ml. (Is the sample representative of a large consignment wit a mean of 130 ml.) and a standard deviation of 10 ml.? Mention the level of significance you use.

18. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are foun to be foggy. What are the probable limits to the percentage of foggy days in the district?

19. Suppose the following ten values represent random observations from a normal parent population:

$$2, 6, 7, 9, 5, 1, 0, 3, 5, 4.$$

Construct a 99 per cent confidence interval for the mean of the parent population.

20. A survey result of 1600 Playboy readers indicates that 44% finished at least three years of college. Se 98% confidence limits on the true proportion of all Playboy readers with this background.

21. (a) What are the alternative approaches of determining a sample size? Explain.

    (b) If we want to draw a simple random sample from a population of 4000 items, how large a sample do we need to draw if we desire to estimate the per cent defective within 2 % of the true value wit 95.45% probability.

22. (a) The following information is given:

    (i) Universe with N=10,000.

    (ii) Variance of weight of the cereal containers on the basis of past records = 8 kg. Determine the size of the sample for estimating the true weight of the containers if the estimate should be within 0.4 kg. of the true average weight with 95% probability.

    (b) What would be the size of the sample if infinite universe is assumed in question number 22 (a) above?

23. Annual incomes of 900 salesmen employed by Hi-Fi Corporation is known to be approximately normally distributed. If the Corporation wants to be 95% confident that the true mean of this year's salesmen's income does not differ by more than 2% of the last year's mean income of ₹ 12,000, what sample size would be required assuming the population standard deviation to be ₹ 1500?

. Mr. Alok is a purchasing agent of electronic calculators. He is interested in determining at a confidence level of 95% what proportion (within plus or minus 4%), is defective. Conservatively, how many calculators should be tested to find the proportion defective?

(***Hint:*** If he tests conservatively, then $p = .5$ and $q = .5$).

5. A team of medico research experts feels confident that a new drug they have developed will cure about 80% of the patients. How large should the sample size be for the team to be 98% certain that the sample proportion of cure is within plus and minus 2% of the proportion of all cases that the drug will cure?

6. Mr. Kishore wants to determine the average time required to complete a job with which he is concerned. As per the last studies, the population standard deviation is 8 days. How large should the sample be so that Mr. Kishore may be 99% confident that the sample average may remain within ± 2 days of the average?

7. I. If $x_1, x_2, x_3, x_4, x_5, x_6$ is a random sample from ***N(0, 1)***, the sampling distribution of $\sum_{i=1}^{6} x_i^2$ is

   (a) *F* with (6, 1) degree of freedom.
   (b) Chi Square with 6 degree of freedom.
   (c) Chi Square with 5 degree of freedom.
   (d) *t* with 5 degree of freedom.

II. If $x_1, x_2, x_3, x_4, x_5, x_6$ is a random sample from ***N(0, s)***, the sampling distribution of $\frac{\overline{x}}{s/\sqrt{6}}$ is

   (a) $N(0, 1)$
   (b) $N(0, s)$
   (c) *t* with 5 degree of freedom
   (d) *t* with 6 degree of freedom

III. If x_1, x_2, x_3, x_4 is a random sample from ***N(0, s)***, the sampling distribution of $\frac{3x_1^2}{(x_2^2 + x_3^2 + x_4^2)}$ is

   (a) Chi square with 4 degree of freedom
   (b) F with (3, 1) degree of freedom
   (c) F with (1, 3) degree of freedom
   (d) $N(0, 1)$

IV. According to the central limit theorem, if a sample of size 81 is drawn from a population with a standard deviation of 72, the standard deviation of the distribution of the sample means would equal ________.

   (a) 8
   (b) 9
   (c) 72
   (d) 0.875

V. In order to evaluate the service level provided to walk-in customers in a multinational bank, th manager records the waiting time of 50 randomly selected walk-in-customers and determined th their mean waiting time was 15 minutes. Assume that the population standard deviation is 4 minute The 90% confidence interval for the population mean of waiting times is

(a) $15 \mp z_{0.05} \times 4/\sqrt{50}$

(b) $15 \mp z_{0.10} \times 4/\sqrt{50}$

(c) $15 \mp t_{0.05} \times 4/\sqrt{50}$

(d) $15 \mp t_{0.10} \times 4/\sqrt{50}$

VI. In order to assess the overall risk of stocks in an economic sector, a fund manager studies a sampl 14 utility stocks from a portfolio of 250 common stocks. In the sample, the mean annualized retur was 14% and that the variance was 3%. Assume that annualized returns are normally distributed The 90% confidence interval for the population variance of annualized returns is ________.

(a) $\left(\dfrac{13\times 3}{\chi^2_{(0.05)}}, \dfrac{13\times 3}{\chi^2_{(0.95)}}\right)$

(b) $\left(\dfrac{13\times 3}{z_{0.05}}, \dfrac{13\times 3}{1+z_{0.05})}\right)$

(c) $\left(\dfrac{3}{1+z_{0.05}\times\sqrt{2/13}}, \dfrac{3}{1-z_{0.05}\times\sqrt{2/13}}\right)$

(d) $14 \mp \chi^2_{(0.05)} \times \sqrt{3}/\sqrt{14}$

VII. A researcher wants to determine the sample size necessary to adequately conduct a study to estimate the population mean to within 5 points. The range of population values is 80 and the researcher plans to use a 90% level of confidence. The sample size should be at least ________ ($Z_{0.05} = 1.645$).

(a) 19

(b) 62

(c) 5

(d) 23

10

Chapter

Testing of Hypothesis

ypothesis is usually considered as the principal instrument in research. Its main function is to ggest new experiments and observations. In fact, many experiments are carried out with the eliberate object of testing hypotheses. Decision-makers often face situations wherein they are terested in testing hypotheses on the basis of available information and then take decisions on the asis of such testing. In social science, where direct knowledge of population parameter(s) is rare, ypothesis testing is the often used strategy for deciding whether a sample data offer such support or a hypothesis that generalisation can be made. Thus hypothesis testing enables us to make probability atements about population parameter(s). The hypothesis may not be proved absolutely, but in practice is accepted if it has withstood a critical testing. Before we explain how hypotheses are tested rough different tests meant for the purpose, it will be appropriate to explain clearly the meaning of hypothesis and the related concepts for better understanding of the hypothesis testing techniques.

10.1 WHAT IS A HYPOTHESIS?

rdinarily, when one talks about hypothesis, one simply means a mere assumption or some supposition be proved or disproved. But for a researcher hypothesis is a formal question that he intends to esolve. Thus a hypothesis may be defined as a proposition or a set of proposition set forth as an xplanation for the occurrence of some specified group of phenomena either asserted merely as a rovisional conjecture to guide some investigation or accepted as highly probable in the light of stablished facts. Quite often a research hypothesis is a predictive statement, capable of being tested y scientific methods, that relates an independent variable to some dependent variables. For example, onsider statements like the following ones:

"Students who receive counselling will show a greater increase in creativity than students not eceiving counselling." Or

"The automobile *A* is performing as well as automobile *B*."

hese are hypotheses capable of being objectively verified and tested. Thus, we may conclude that hypothesis states what we are looking for and it is a proposition which can be put to a test to etermine its validity.

Characteristics of hypothesis: Hypothesis must possess the following characteristics:

(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, inferences drawn on its basis cannot be taken as reliable.

(ii) Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, ma a time the research programmes have bogged down. Some prior study may be done researcher in order to make hypothesis a testable one. A hypothesis is testable if ot deductions can be made from it which, in turn, can be confirmed or disproved by observati

(iii) Hypothesis should state relationship between variables, if it happens to be a relatio hypothesis.

(iv) Hypothesis should be limited in scope and must be specific. A researcher must rememb that narrower hypotheses are generally more testable and he should develop such hypothes

(v) Hypothesis should be stated as far as possible in most simple terms so that the same easily understandable by all concerned. But one must remember that simplicity of hypothe has nothing to do with its significance.

(vi) Hypothesis should be consistent with most known facts i.e., it must be consistent with substantial body of established facts. In other words, it should be one which judges acce as being the most likely.

(vii) Hypothesis should be amenable to testing within a reasonable time. One should not u even an excellent hypothesis, if the same cannot be tested in reasonable time for o cannot spend a life-time collecting data to test it.

(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This mea that by using the hypothesis plus other known and accepted generalizations, one should b able to deduce the original problem condition. Thus hypothesis must actually explain wh it claims to explain; it should have empirical reference.

10.2 BASIC CONCEPTS CONCERNING TESTING OF HYPOTHESIS

Basic concepts in the context of testing of hypotheses need to be explained.

10.2.1 Null Hypothesis and Alternative Hypothesis

In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. I we are to compare method A with method B about its superiority and if we proceed on the assumptio that both methods are equally good, then this assumption is termed as the null hypothesis. As agains this, we may think that the method A is superior or the method B is inferior, we are then stating wha is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternativ hypothesis as H_1. Suppose we want to test the hypothesis that the population mean (μ) is equal t the hypothesised mean $(\mu_0) = 100$. Then we would say that the null hypothesis is that the populatio mean is equal to the hypothesised mean 100 and symbolically we can express as:

$$H_0 : \mu = \mu_0 = 100$$

If our sample results do not support this null hypothesis, we should conclude that something else rue. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other rds, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we cept H_0, then we are rejecting H_1 and if we reject H_0, then we are accepting H_1. For $_0 : \mu = \mu_0 = 100$, we may consider three possible alternative hypotheses as follows:

Table 10.1

lternative hypothesis	*To be read as follows*
$H_1 : \mu \neq \mu_0$	(The alternative hypothesis is that the population mean is not equal to 100 i.e., it may be more or less than 100)
$H_1 : \mu > \mu_0$	(The alternative hypothesis is that the population mean is greater than 100)
$H_1 : \mu < \mu_0$	(The alternative hypothesis is that the population mean is less than 100)

he null hypothesis and the alternative hypothesis are chosen before the sample is drawn (the researcher ust avoid the error of deriving hypotheses from the data that he collects and then testing the ypotheses from the same data). Alternative and null hypotheses are the statements about unknown opulation parameters. In null hypothesis, we should always have 'equal to' sign. Null hypothesis is e specific statement about the parameter, e.g., $H_0 : \mu = 50$.

Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are ying to reject, and alternative hypothesis represents all other possibilities.

0.2.2 Type I and Type II Errors

n the context of testing of hypotheses, there are basically two types of errors we can make. We nay reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is nown as Type I error and the latter as Type II error. In other words, Type I error means rejection of ypothesis which should have been accepted and Type II error means accepting the hypothesis vhich should have been rejected. All the possibilities in decision making using hypothesis testing are iven in the tabular form as below:

Table 10.2

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Accept H_0	No Error	Type II Error
	Probability $=1-\alpha$	Probability $= \beta$
Reject H_0	Type I Error	No Error
	Probability $= \alpha$	Probability $= 1-\beta$

Also, the size of Type I Error is given by the probability of Type I error, P (Reject H_0| H_0) whi is denoted by α. In statistical quality control. Similarly, the size of Type II Error is given by t probability of Type II error, P (Accept H_0| H_1) which is denoted by β. In statistical quality control is called as producer's risk and β is called as consumer's risk.

The probability of Type I error is usually determined in advance and is understood as the level significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there a about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximu probability of committing Type I error would only be 0.01.

But with a fixed sample size, *n*, when we try to reduce Type I error, the probability of committin Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-o between two types of errors which means that the probability of making one type of error can on be reduced if we are willing to increase the probability of making the other type of error. To deal wi this trade-off in business situations, decision-makers decide the appropriate level of Type I error b examining the costs or penalties attached to both types of errors.

10.2.3 Level of Significance

As discussed, level of significance (α) is the probability of Type I error. This is very importan concept in the context of hypothesis testing. It is always some percentage (usually 5%) which shoul be chosen wit great care, thought and reason. In case we take the significance level at 5 per cen then this implies that H_0 will be rejected when the sampling result (i.e., observed evidence) has a les than 0.05 probability of occurring if H_0 is true. In other words, the 5 per cent level of significanc means that researcher is willing to take as much as a 5 per cent risk of rejecting the null hypothesi when it (H_0) happens to be true. Thus the significance level is the maximum value of the probabilit of rejecting H_0 when it is true and is usually determined in advance before testing the hypothesis.

10.2.4 Two-Tailed and One-Tailed Tests

As we discussed we test three types of hypotheses given by

(i) $H_0 : \mu = \mu_0$ Against $H_1 : \mu \neq \mu_0$

(ii) $H_0: \mu = \mu_0$ Against $H_1: \mu > \mu_0$ or $H_0: \mu \leq \mu_0$ Against $H_1: \mu > \mu_0$

(iii) $H_0: \mu = \mu_0$ Against $H_1: \mu < \mu_0$ or $H_0: \mu \geq \mu_0$ Against $H_1: \mu < \mu_0$

Based on the sign in alternative hypothesis ($\neq$, >, or <), we have three different tests. When we e '$\neq$' sign in alternative hypothesis, we have two-tailed test; when we have '>' sign in alternative othesis, we have right-tailed test; and for '<' sign in alternative hypothesis, we have left-tailed .

0.3 TESTING THE HYPOTHESIS

ven a hypothesis H_0 and an alternative hypothesis H_1, we make a rule which is known as decision e according to which we accept H_0 (i.e., reject H_1) or reject H_0 (i.e. accept H_1). For example, pose we want to examine that the mean age of the people in a city is 40 years. In order to conduct hypothesis testing, we need to be a bit more specific if we wish to examine that

(i) the mean age of the people in a city is 40 years or not; or

(ii) the mean age of the people in a city is 40 years or higher; or

(iii) the mean age of the people in a city is 40 years or lower.

For testing above claims, we first setup the hypotheses which are given as below:

(i) $H_0: \mu = 40$ Against $H_1: \mu \neq 40$ (for claim (i))

(ii) $H_0: \mu = 40$ Against $H_1: \mu > 40$ (for claim (ii))

(iii) $H_0: \mu = 40$ Against $H_1: \mu < 40$ (for claim (iii))

Now we draw a probabilistic (or random) sample of a size from the aforesaid population. Size of e sample is already known. Since we are testing the claim about population mean, we obtain mple mean as sample mean is a "good" estimate of population mean. Suppose sample mean mes to be 20 years. This is significantly lower than the claimed mean population age 40 years. If e claim (***H***$_0$) is true, the probability of getting such a different sample mean would be very small. etting a sample mean of 20 is very unlikely. So, when we get sample mean as 20, we do not believe n the claim (***H***$_0$). If the sample mean is close to the assumed population mean, ***H***$_0$ is accepted. If e sample mean is far-off from the assumed population mean, ***H***$_0$ is rejected. How far is "far nough" to reject ***H***$_0$? The concept of critical value is used to decide on this (to be discussed later).

Also, in a hypothesis test, we initially assume that the null hypothesis is true and we proceed to y to reject null hypothesis using the sample. In case, when we cannot reject the null hypothesis it nly means that sample has insufficient information to reject null hypothesis at given level of gnificance. It does not mean that the parametric statement under alternative hypothesis is true. herefore, whenever we say that the null hypothesis is accepted, it only means that null hypothesis annot be rejected as there is no statistical evidence against it.

10.4 TEST STATISTIC AND CRITICAL REGION

fter setting up the hypothesis we obtain the required test statistic using the sample observations nd the hypothetical value of the parameter under consideration. The formulae of test statistic under different requirements and conditions are discussed later. Since test statistic is calculated f sample observations, it is a statistic and has some sampling distribution. In case of large samples, distribution of test statistic becomes normal distribution, while in case of small samples we use o sampling distributions, e.g., *t*-distribution, *F*-distribution, and chi-square distribution.

The distribution of test statistic is used in order to reject or accept the null hypothesis. Crit value divides the area under probability curve of the distribution of test statistic into two region critical (or rejection) region and acceptance region. Size of critical region is given by level of significa α, while the size of acceptance region is $(1-\alpha)$. When the statistical outcome (the value of statistic) falls into the critical region, H_0 is rejected. When the statistical outcome falls into accepta region, H_0 is accepted.

10.5 CRITICAL VALUE AND DECISION RULE

Critical value is obtained using the sampling distribution of test statistic and prefixed level significance. As discussed, critical value divides the area under probability curve of the distributi of test statistic into critical region and acceptance region. Also, the size of critical region is the sa as level of significance α.

In case of two-tailed test, when we wish to test H_0: $\mu = \mu_0$ against $H_1: \mu \neq \mu_0$, we have tw critical values as the critical region is divided into two equal parts on the two ends of probabili curve. Size of each part is $\alpha/2$.

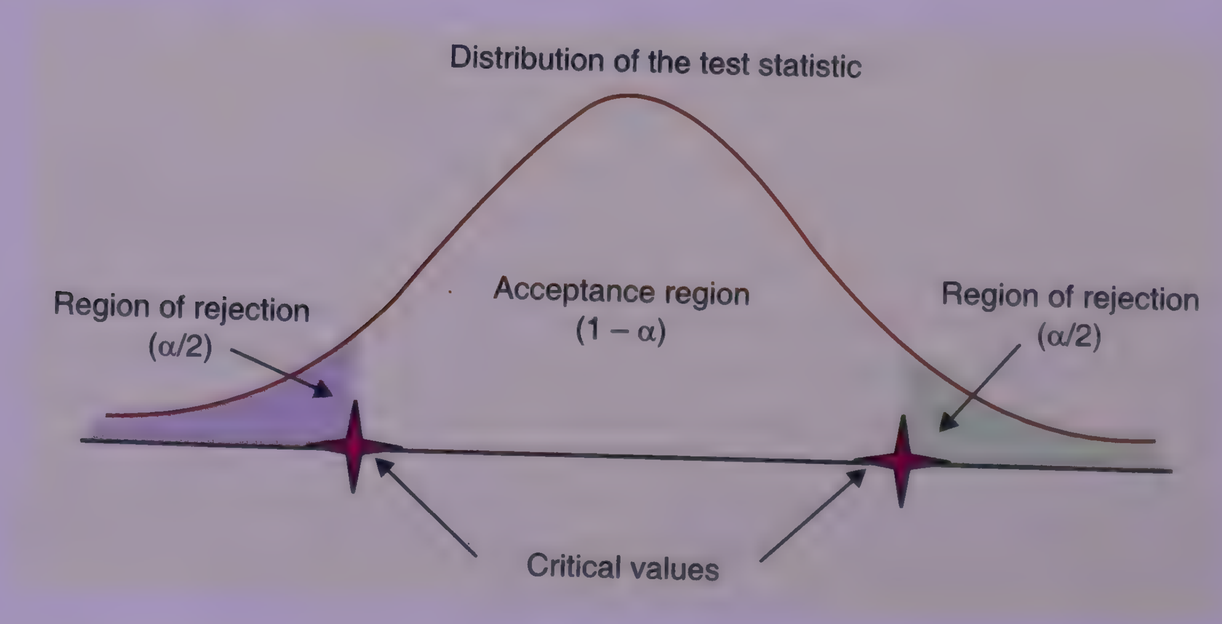


Fig. 10.1

In the figure, we show critical region, acceptance region and critical values. When the distribution o test statistic is N(0, 1) and the level of significance is 5% or $\alpha = 0.05$, the critical values are –1.96 an 1.96. It means that the null hypothesis is rejected when the value of test statistic is either lower tha –1.96 or greater than 1.96.

For a right tailed test, when we wish to test H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$, the critical value is o the right side end of the probability curve as shown in Figure.

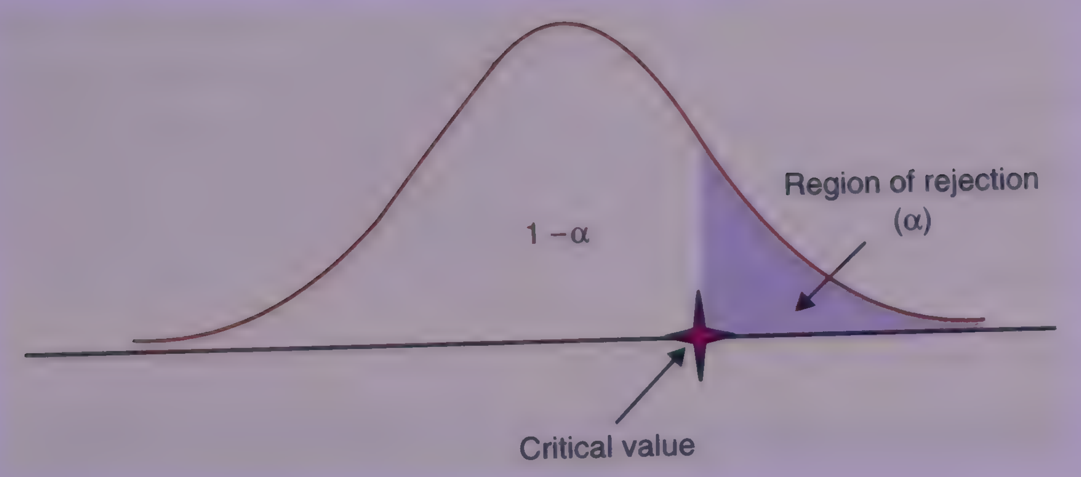


Fig. 10.2

When the distribution of test statistic is N(0, 1) and the level of significance is 5% or $\alpha = 0.05$, the tical value is 1.645. The null hypothesis is rejected when the value of test statistic is smaller than 45.

For a right tailed test, when we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$, the critical value is the right side end of the probability curve as shown in Figure.

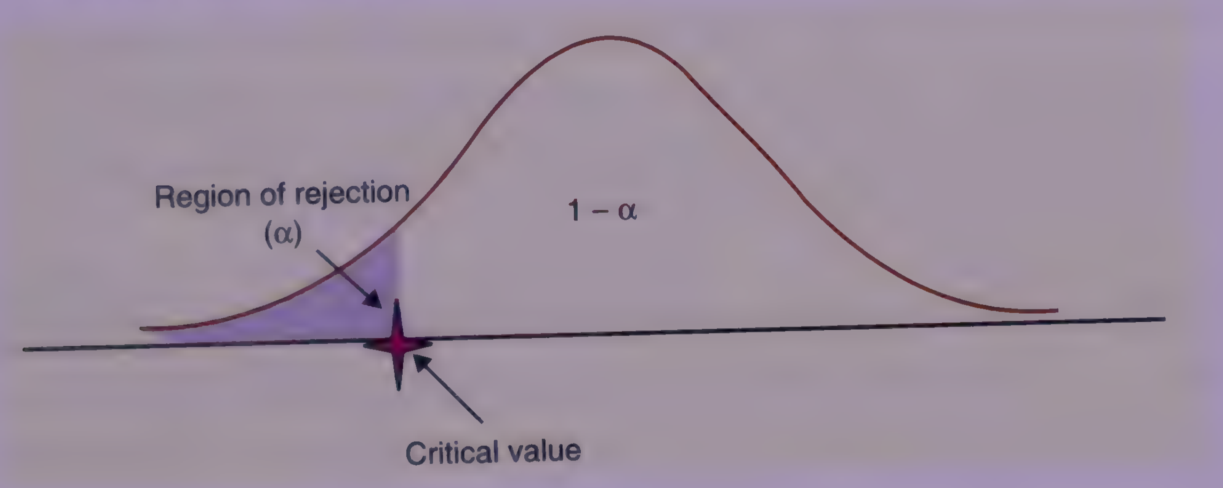


Fig. 10.3

When the distribution of test statistic is N(0, 1) and the level of significance is 5% or $\alpha = 0.05$, the ritical value is –1.645. The null hypothesis is rejected when the value of test statistic is smaller than 1.645.

10.6 PROCEDURE FOR HYPOTHESIS TESTING

o test a hypothesis means to tell (on the basis of the data the researcher has collected) whether or ot the hypothesis seems to be valid. In hypothesis testing the main question is: whether to reject the ull hypothesis or not to reject the null hypothesis? Procedure for hypothesis testing refers to all those teps that we undertake for making a choice between the two actions i.e., rejection and acceptance f a null hypothesis. The various steps involved in hypothesis testing are stated below:

(i) *Setting up the hypotheses:* The consists in making a formal statement of the null hypoth (H_0) and also of the alternative hypothesis (H_1). This means that hypotheses should be cle stated, considering the nature of the research problem. For instance, Mr. Mohan of the C Engineering Department wants to test the load bearing capacity of an old bridge which m be more than 10 tons, in that case he can state his hypotheses as under:

Null hypothesis $H_0 : \mu = 10$ tons

Alternative hypothesis $H_1 : \mu > 10$ tons

Take another example. The average score in an aptitude test administered at the natio level is 80. To evaluate a state's education system, the average score of 100 of the stat students selected on random basis was 75. The state wants to know if there is a signific difference between the local scores and the national scores. In such a situation the hypothe may be stated as under:

Null hypothesis $H_0 : \mu = 80$

Alternative hypothesis $H_1 : \mu \neq 80$

The formulation of hypotheses is an important step which must be accomplished with d care in accordance with the object and nature of the problem under consideration. It al indicates whether we should use a one-tailed test or a two-tailed test. If H_1 is of the ty greater than (or of the type lesser than), we use a one-tailed test, but when H_1 is of the ty "whether greater or smaller" then we use a two-tailed test.

(ii) *Selecting a significance level:* The hypotheses are tested on a pre-determined level significance and as such the same should be specified. Generally, in practice, either 5 level or 1% level is adopted for the purpose. The factors that affect the level of significan are: (a) the magnitude of the difference between sample means; (b) the size of the sample (c) the variability of measurements within samples; and (d) whether the hypothesis directional or non-directional (A directional hypothesis is one which predicts the directi of the difference between, say, means). In brief, the level of significance must be adequat in the context of the purpose and nature of enquiry.

(iii) *Test statistic:* We will be conducting hypothesis tests (or tests of significance) for mea proportion, and variance. The formulae for test statistics and their distributions are discusse later. At this step, we obtain the value of test statistic using the sample observations selecte by the researcher and the hypothetical parametric value stated under null hypothesis.

(iv) *Critical Value:* Using the distribution of test statistic, level of significance α, and the typ of test (two-tailed, right-tailed, or left-tailed), we obtain the critical value.

(v) *Decision:* Comparing the value of test statistic and critical value, we make the decisio about rejecting or not rejecting the null hypothesis.

We reject the null hypothesis when

(a) value of test statistic < lower critical value or value of test statistic > upper critical value, in case of a two-tailed test;

(b) value of test statistic > critical value, in case of a right-tailed test;

(c) value of test statistic < critical value, in case of a left-tailed test.

Otherwise, we cannot reject the null hypothesis as we do not have enough statistical evidence inst it.

0.7 HYPOTHESIS TESTING FOR MEAN

esting of mean, we test on the basis of a random if the mean of a population mean is the same as hypothesized value or different (or higher, or smaller). In technical terms, we may like test any he following hypotheses:

(i) $H_0: \mu=\mu_0$ Against $H_1: \mu\neq\mu_0$.

(ii) $H_0: \mu=\mu_0$ Against $H_1: \mu>\mu_0$ or $H_0: \mu\leq\mu_0$ Against $H_1: \mu>\mu_0$.

(iii) $H_0: \mu=\mu_0$ Against $H_1: \mu<\mu_0$ or $H_0: \mu\geq\mu_0$ Against $H_1: \mu<\mu_0$.

Here μ_0 is the hypothetical population mean. The test is different for the situation when population ndard deviation (σ) is known than that is used in the situation when σ is not known.

se 1: σ is known

this case the population should be normally distributed or sample should be large enough (in actice, $n\geq 30$). The test statistic is

$$Z_c=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

ere, $\bar{x}$ = sample mean, σ = known population standard deviation, and n = sample size.

The distribution of this test statistic under the above mentioned conditions is N(0, 1). We below ate an example to illustrate the method.

xample 10.1: A survey, done 10 years ago, of CPA's (Certified Public Accountant) in the U.S. und that their average salary was $74,914. An accounting researcher would like to test whether er the years this average has increased or not? A sample of 112 CPAs produced a mean salary of 8,695. Also given that the population standard deviation σ = $14,530.

lution: We first setup the test hypotheses. Here we wish to test:

$H_0: \mu\leq 74914$ against $H_1: \mu> 74914$, (right tailed test)

Now, let us fix the level of significance as 5%. There is no universally accepted level of significance. esearcher need not to fix the size of Type II error in advance. However, he/she should try to have optimum size of Type I error so that the size of Type II error is also not very large.

As we can notice that population standard deviation σ = \$14,530 is known and sample $n = 112$, which is large enough. Therefore, the test statistic is

$$Z_c = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{78695 - 74914}{14530/\sqrt{112}} = 2.7539$$

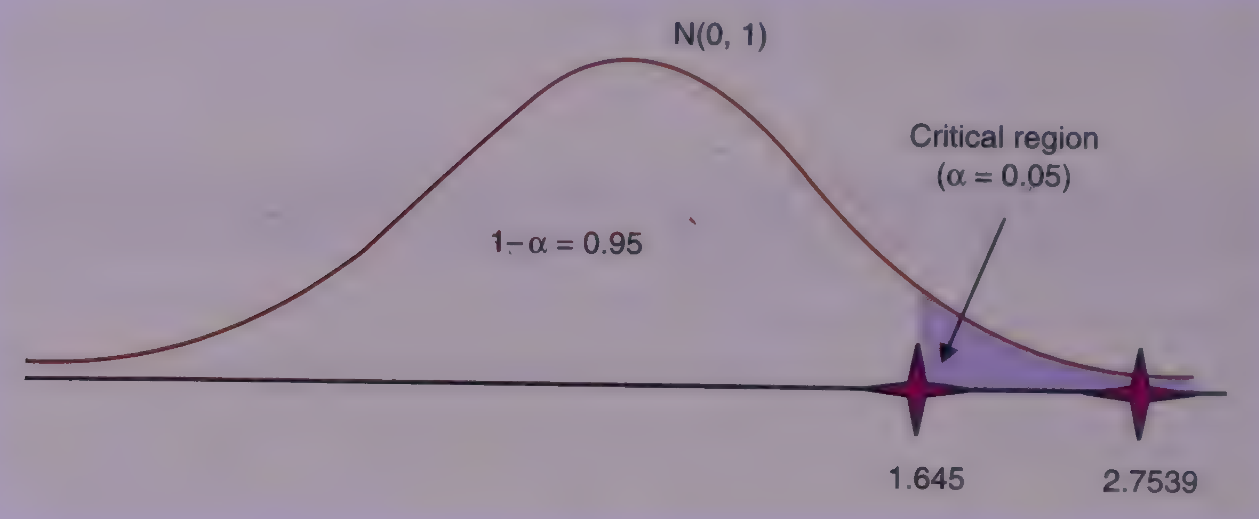


Fig. 10.4

Distribution of test statistic is N(0, 1). So critical value for right tailed test and for 5% level significance is 1.645. Since, computed value > critical value at 5% level of significance, we rej H_0 at 5% level of significance in favor of H_1 and conclude that average salary of CPAs in the U has increased.

Case 2: σ is not known

In this case the population should be normally distributed. When population is not normal we requ a large sample as in case of a sufficiently large sample standard deviation and population standa deviation will almost be the same and we can proceed using the method meant for σ known ca Thus, this method is generally used for small sample drawn from a normal population. The te statistic is

$$T_c = \frac{\overline{x} - \mu_0}{s_1/\sqrt{n}}$$

where, $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$ is sample standard deviation. The distribution of this test statistic is student's *t*-distribution with $(n-1)$ degree of freedom. Critical values of $t_{(n-1)}$ distribution can obtained from the *t* table for given d.f. and significance level. Student's *t*-distribution is symmetr about *y*-axis like N(0, 1) distribution. The plot of *t*-distribution is almost the same as of N(0, distribution. The only difference is, *t*-distribution's probability curve has more spread and less pea This test is also called as *t* test for single sample. To illustrate the method, consider the followi example:

Example 10.2: A sample of 25 people is taken. The length of time to prepare dinner is recorded minutes, as given below:

44.0	51.9	49.7	40.0	55.5	33.0	43.4	41.3	45.2	40.7	41.1	49.1
30.9	45.2	55.3	52.1	55.1	38.8	43.1	39.2	58.6	49.8	43.2	47.9
46.6											

Is there any evidence that the population mean time to prepare dinner is less than 48 minutes? a level of significance of 0.05.

ution: We wish to test

$$H_0 : \mu \geq 48 \text{ against } H_1 : \mu < 48 \text{. (left tailed test)}$$

Population standard deviation is not known. Using the given data, we calculate $n=25$, $\overline{x}=45.628$, 6.9587. Test statistic:

$$T_c = \frac{\overline{x} - \mu_0}{s_1 / \sqrt{n}} = -1.70435$$

The test statistic follows Student's t-distribution with $n-1=24$ degree of freedom. For a left led test and for level of significance of 0.05, the critical value using t-distribution with $n-1=24$ gree of freedom (from Table 2 in Appendix) is –2.064.

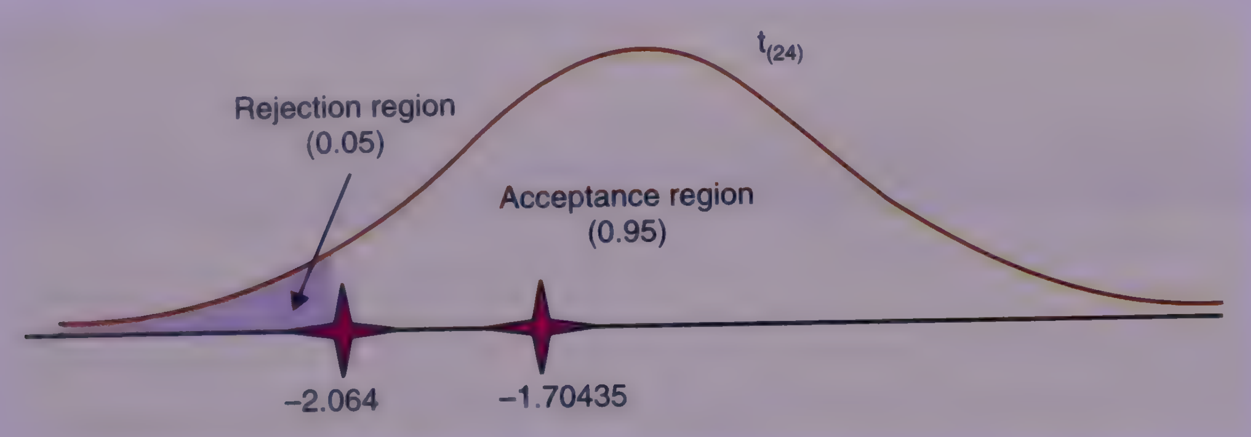


Fig. 10.5

Since calculated value of test statistic > critical value, we cannot reject the null hypothesis at ven level of significance of 0.05.

You can notice here that sample mean is 45.628 which is less than 48. However using the test of pothesis we do not have sufficient statistical evidence to reject null hypothesis stating that the pulation mean is greater than or equal to 48. Thus we do not have enough statistical evidence at ven level of significance to favor the alternative hypothesis stating that the population mean is naller than 48. The reason of this is high variance in sample observations.

Note: Note that in case of finite population (when $n / N > 0.05$), you should use finite opulation correction (*fpc*). In such a case, we multiply the denominator of both of the test statistic given by Z_c and T_c, by $fpc = \sqrt{\frac{N-n}{N-1}}$. Rests of the steps are the same.

10.8 HYPOTHESIS TESTING FOR PROPORTION

In this test the population is divided into two mutually exclusive and exhaustive classes basec certain attribute—one class having that attribute and another class not having that attribute. parameter of interest here is the population proportion or the proportion of population having attribute, e.g., proportion of female students in a college or the proportion of defective ite manufactured by some manufacturing unit in the whole day. Based on a sample, we test any of following hypotheses:

- $H_0 : \pi \leq \pi_0, H_1 : \pi > \pi_0$ (Right or Upper Tail Test)
- $H_0 : \pi \geq \pi_0, H_1 : \pi < \pi_0$ (Left or Lower Tail Test)
- $H_0 : \pi = \pi_0, H_1 : \pi \neq \pi_0$ (Two Upper Tail Test)

The test statistic for this test is

$$Z_c = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

where, p is the sample proportion. We have seen in the previous chapter that the distribution of t test statistic is N(0, 1) when sample is sufficiently large (in practice, $n \geq 30$ or $n\pi_0 \geq 5$ as well $n\pi_0(1 - \pi_0) \geq 5$. To illustrate this test, consider the following example:

Example 10.3: A marketing company claims that it receives 8% responses from its mailing. seems that the company is giving an exaggerated picture and the said percentage is less. To test tl claim, a random sample of 500 was surveyed with 30 responses. Test the company's claim at t $\alpha = 0.05$ significance level against that the company is giving and exaggerated picture.

Solution: We wish to test:

$$H_0: \pi = 0.08 \text{ against } H_1: \pi < 0.08 \text{ (left tailed test)}$$

From the given data,

Sample proportion $p = 30/500 = 0.06$;

Sample size $n = 500$ (large enough).

Test Statistic:

$$Z_c = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = \frac{0.06 \times 0.08}{\sqrt{0.08 \times 0.92/500}} = -1.64845$$

Level of significance is given as 0.05. The distribution of test statistic is N(0, 1). Therefore tl critical value, obtained using N(0, 1) distribution, is – 1.645.

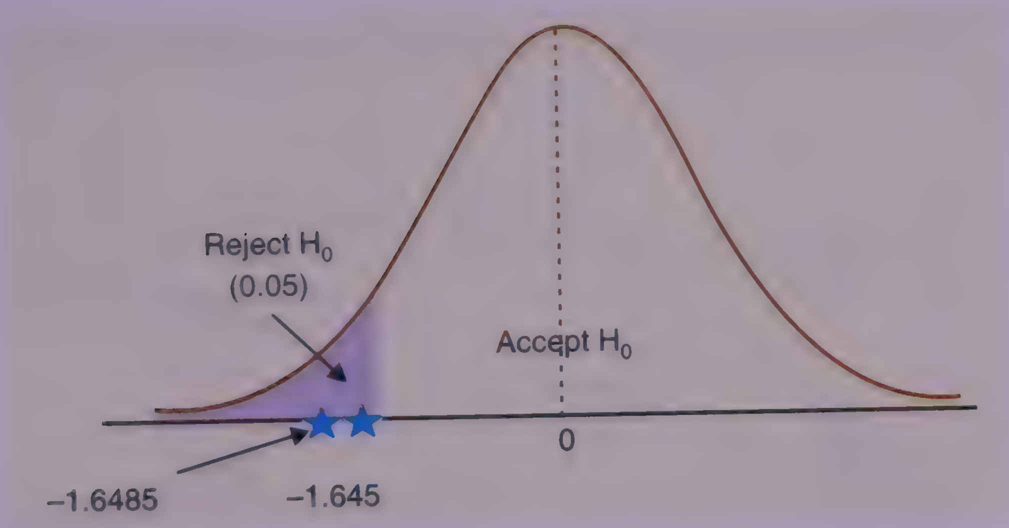


Fig. 10.6

At 5% level of significance the computed value of test statistic falls in the critical region, therefore eject the null hypothesis at 5% level of significance. Note that if we change the level of significance .5% than the critical value will be −1.96 and we cannot reject the null hypothesis. Thus the earcher needs to decide the level of significance carefully. Non rejection of the null hypothesis is sed by unavailability of any statistical evidence against the null hypothesis provided by the selected nple.

.9 HYPOTHESIS TESTING FOR VARIANCE

e we wish to test any of the following hypotheses:

(i) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$ (two tailed test)

(ii) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$ (right tailed test)

(iii) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$ (left tailed test)

The test statistic:

$$\chi_c^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sigma_0^2} = \frac{(n-1)s_1^2}{\sigma_0^2}$$

ere, $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$ is the sample variance and σ_0^2 is the hypothesized value of population iance. We have seen in the previous chapter that the distribution of test statistic χ_c^2 is chi-square tribution with $(n-1)$ degree of freedom provided that the population has normal distribution. The ge of N(0, 1) distribution and *t*-distribution are the same as $(-\infty, \infty)$, while the range of chi- are distribution is $(0, \infty)$. Also unlike N(0, 1) and *t*-distributions, chi-square distribution is not nmetric. When the test statistic follows chi square distribution, critical values are obtained from chi square table (Table 3 in Appendix) for given level of significance and degree of freedom.

For a two tailed test $(H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2)$ the critical region and critical values are displ in the following figure:

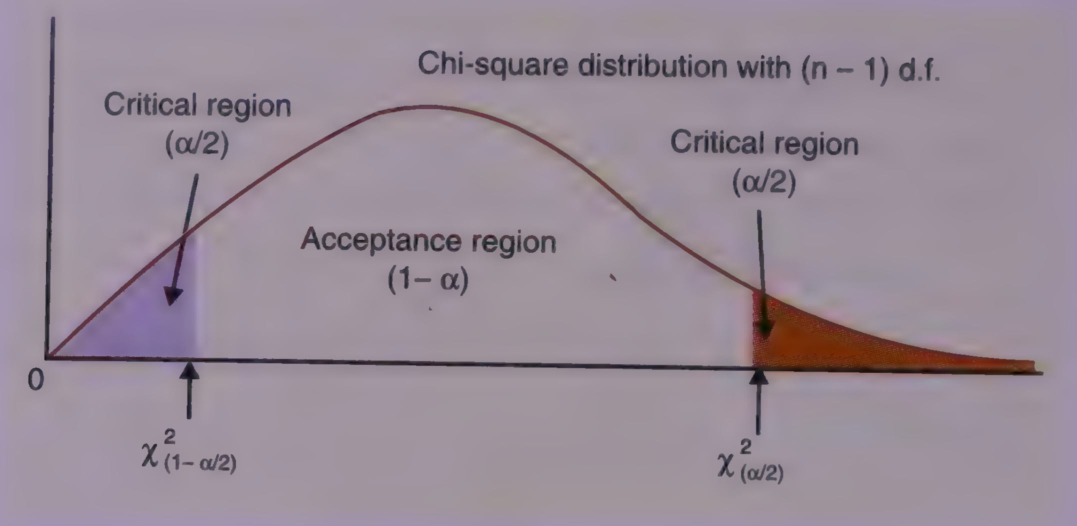


Fig. 10.7

Chi-square distribution is not symmetric. So, we have different notations for critical values. A after the point is mentioned in subscript. The notation $\chi^2_{(1-\alpha/2)}$ means that the area after this poin

∞ below the probability curve is $(1-\alpha/2)$. Similarly, $\chi^2_{(\alpha/2)}$ means that the area after this point til below the probability curve is $\alpha/2$. We reject H_0 in the favor of H_1 at $\alpha \times 100\%$ level, if the compu value of test statistic is smaller than $\chi^2_{(1-\alpha/2)}$ or higher than $\chi^2_{(\alpha/2)}$.

For a right tailed test $(H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2)$ the critical region and critical values are displa in the following Fig. 10.8.

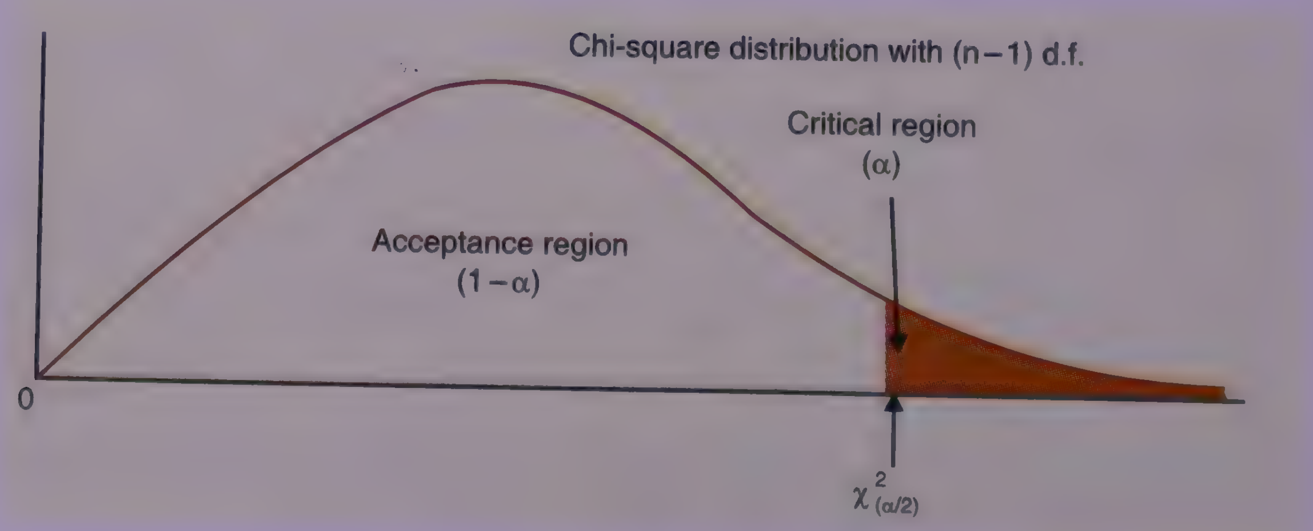


Fig. 10.8

The critical region is on the right tail of the probability curve. The critical value is denoted as . We reject H_0 in the favor of H_1 at $\alpha \times 100\%$ level, if the computed value of test statistic is er than $\chi^2_{(\alpha)}$.

Similarly, for a left tailed test $(H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2)$ the critical region and critical values be understood using the following Fig. 10.9.

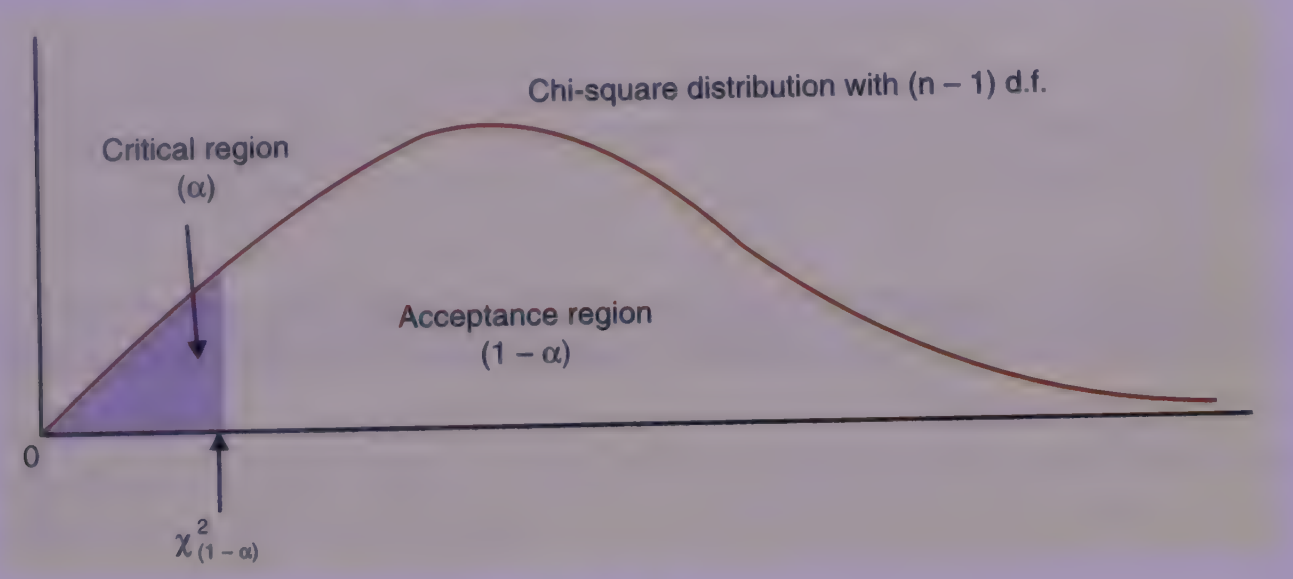


Fig. 10.9

The critical value is denoted as $\chi^2_{(1-\alpha)}$ as the area after this point till infinity below the probability ve is (1–α) (the size of acceptance region). We reject H_0 in the favor of H_1 at $\alpha \times 100\%$ level, if computed value of test statistic is lower than $\chi^2_{(1-\alpha)}$.

The method is illustrated using the following example:

ample 10.4: A pharmaceutical company is considering the purchase of new bottling machines to rease efficiency. The factory currently makes use of machines that fill cough syrup bottles whose ume of medicine has a standard deviation of 1.6 mL. The new machine they are considering was ted on 30 bottles, producing a batch with a standard deviation of 1.2 mL. Does this machine duce a standard deviation less than 1.6 mL? Assuming normal distribution, test at 5% significance el?

ution: Here we wish to test

$$H_0 : \sigma^2 \geq (1.6)^2 \text{ against } H_1 : \sigma^2 < (1.6)^2$$

$$\text{or } H_0 : \sigma^2 = (1.6)^2 \text{ against } H_1 : \sigma^2 < (1.6)^2$$

$$\text{Test statistic: } \chi_c^2 = \frac{(n-1)s_1^2}{\sigma_0^2} = \frac{(30-1)\times 1.2\times 1.2}{1.6\times 1.6} = 16.3125$$

The distribution of test statistic is chi-square with d.f. = 29. It is a left tailed test with lev significance as 5%. Therefore critical value, using Table 3 in Appendix, is 17.708. Since comp value of test statistic is smaller than the critical value, we reject the null hypothesis at 5% lev significance and conclude that the new machine produces a standard deviation less than 1.6 m

Note: Table 3 has the critical values for chi-square distribution with degree of freedom 1 to

For larger samples N(0, 1) distribution can be used. After obtaining the value of test statistic apply following formula:

$$Z_c = \sqrt{2\chi^2_\alpha} - \sqrt{2n-1}$$

Distribution of this Z_c is N(0, 1). The computed value of Z_c can be compared with the criti values obtained using N(0, 1) distribution.

10.10 HYPOTHESIS TESTING FOR DIFFERENCE OF TWO MEAN

In Section 10.8 the last chapter, we had one sample from a population. On the basis of that sam we examined the validity of hypothetical statements about unknown population mean. Here we H be having two samples, may be drawn from different populations, and based on these two sam we compare the population means.

We have a sample $x_1, x_2, ..., x_n$ drawn from a population with population mean μ_x and popula variance σ_x^2. Another sample $y_1, y_2, ..., y_m$ is drawn from a population with population mean μ_y population variance σ_y^2. We wish to test any of the following hypothesis:

(i) $H_0: \mu_x = \mu_y$ Against $H_1: \mu_x \neq \mu_y$

(ii) $H_0: \mu_x = \mu_y$ Against $H_1: \mu_x > \mu_y$ or $H_0: \mu_x \leq \mu_y$ Against $H_1: \mu_x > \mu_y$

(iii) $H_0: \mu_x = \mu_y$ Against $H_1: \mu_x < \mu_y$ or $H_0: \mu_x \geq \mu_y$ Against $H_1: \mu_x < \mu_y$

Under different conditions we follow different methods explained as under:

Case 1: Both samples are independent of each other and both population variances σ_x^2 and σ_y^2 known.

Let $\bar{x}$ and $\bar{y}$ be the sample means. Since we wish to compare the population so parameter of interest is $(\mu_x \times \mu_y)$. A feasible estimate of this unknown parameter is $(\bar{x} \times$

From central limit theorem, for sufficiently large sample, the distribution of $\frac{T - E(T)}{S.E.(T)}$ N(0, 1), where T is a sample statistic, $E(T)$ is the expected value or mean of T and $S.E.$ is the standard error of T. Using the sampling distribution of mean, we can write th

$-\overline{y})=\mu_x \times \mu_y$ and $S.E.(\overline{x}-\overline{y})=\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}$. Since the value of parameter $(\mu_x - \mu_y)$ under hypothesis is zero, the test statistic is

$$Z_c=\frac{(\overline{x}-\overline{y})-0}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}=\frac{\overline{x}-\overline{y}}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}$$

When both populations are normal or both sample sizes are large enough, the distribution of this statistic Z_c is N(0, 1). This test is also called as Z test: two samples for means. The application of method is explained through following example:

mple 10.5: In a survey of buying habits, 400 women shoppers are chosen at random in a super ket 'A'. Their average weekly food expenditure is ₹ 250 with a standard deviation of). For 400 women shoppers chosen at random in some other super market 'B', the average kly food expenditure is ₹ 220 with a standard deviation of ₹ 55. Do these two populations have ilar shopping habits? Is the average weekly food expenditure of two populations of shoppers al? Test at 5% level of significance.

ution: We wish to test

$H_0:\mu_x=\mu_y$ Against $H_1:\mu_x\neq\mu_y$

Both sample sizes are very large. For very large sample, sample variance and population variance almost the same. So, in this example, we consider as population variances be known.

Test Statistic:

$$Z_c=\frac{\overline{x}-\overline{y}}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}=\frac{250-220}{\sqrt{\frac{40\times 40}{400}+\frac{55\times 55}{400}}}=8.822575$$

For a two tailed test, at 5% level of significance, the critical values using N(0, 1) distribution are 96 and 1.96. Since the computed value of test statistic is larger than 1.96, we reject the null othesis at 5% level and conclude that two populations have different shopping habits.

ote: In case we wish to test $H_0:\mu_x=\mu_y+a$ Against $H_1:\mu_x\neq\mu_y+a$ where a is a constant,

e can modify the formula of test statistic as $Z_c=\frac{\overline{x}-\overline{y}-a}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}$.

se 2: Both samples are independent of each other and both population variances σ_x^2 and σ_y^2 are

known.

When both population variances are unknown, we use Student's t distribution which requires conditions:

(i) Both populations have normal distribution, and

(ii) Both population variances are equal, i.e., $\sigma_x^2 = \sigma_y^2$.

Unknown variances are estimated by the pooled sample variance given by

$$S^2 = \frac{1}{m+n-2}\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{j=1}^{m} (y_i - \bar{y})^2\right].$$

The test statistic is $T_c = \dfrac{\bar{x} - \bar{y}}{S\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$

Distribution of T_c is Student's t-distribution with degree of freedom $= m + n - 2$. This test is called as t test for independent samples.

Example 10.6: You are a financial analyst for a brokerage firm. Is there a difference in divid yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number of stocks	21	25
Sample mean	3.27	2.53
Sample std. dev.	1.30	1.16

Assuming both populations are approximately normal with equal variances, is there a differe in average yield ($\alpha = 0.05$)?

Solution: We wish to test:

$$H_0 : \mu_x = \mu_y \quad \text{against} \quad H_1 : \mu_x \neq \mu_y$$

Pooled variance: $S^2 = \dfrac{1}{21+25-2}[20 \times 1.30 \times 1.30 + 24 \times 1.16 \times 1.16] = 1.5021$

Test Statistic:

$$T_c = \frac{3.27 - 2.53}{\sqrt{1.5021} \times \sqrt{\dfrac{1}{21} + \dfrac{1}{25}}} = 2.040$$

Test statistic has t-distribution with 21 + 25 – 2 = 44 degree of freedom. Critical values t-distribution for more than 30 degree of freedom can be approximated using N(0, 1) distributio Thus, the critical values for this two tailed test at 5% level of significance are –1.96 and 1. Computed value of test statistic is 2.040 which is larger than 1.96, so we reject the null hypothesis 5% level.

Using Excel the critical values for t distribution with 44 d.f. are –2.0154 and 2.0154. Comput value of test statistic is 2.040 which is larger than 2.0154, so we reject the null hypothesis at 5% lev

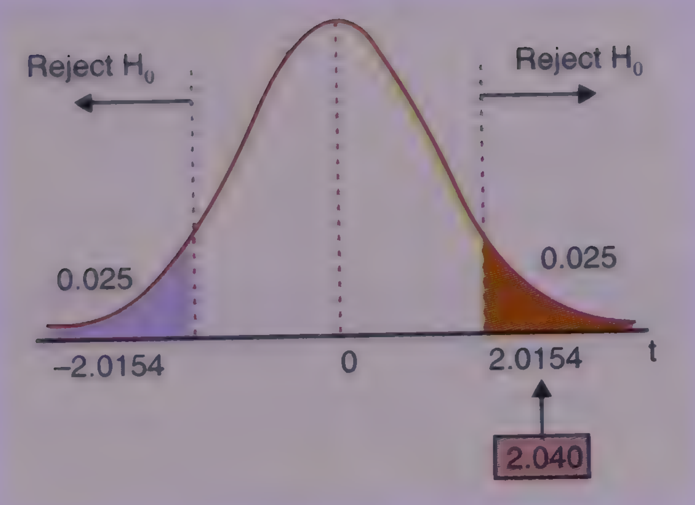


Fig. 10.10

e 3: Both samples are related.

The two samples are not independent but related in sense that the observations are paired. For mple, the height of a group of 15 children before and after they are fed a certain nutritional food duct for two months make two samples which are related and observations are paired. In this mple, we may be interested in testing the effect of the nutritional food product on the height of dren by comparing 'before' and 'after' heights.

Clearly both sample sizes are the same, i.e., $n = m$. We first obtain the difference $= x_i - d_i (i = 1, 2, ..., n)$. Using the differences we obtain the mean $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$ and sample ndard deviation $s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (d_i \times \bar{d})^2}$. The test statistic is

$$T_c = \frac{\bar{d}}{s_d \sqrt{n}}$$

This T_c has t-distribution with $n-1$ d.f.

ample 10.7: You send your salespeople to a "customer service" training workshop. Has the ning made a difference in the number of complaints? You collect the following data:

Table 10.3

Salesperson	*Number of complaints*	
	Before	*After*
A	6	4
B	20	6
C	3	2
D	0	0
E	4	0

Conduct the test assuming normal distribution at 1% level of significance.

Solution: We wish to test:

$$H_0 : \mu_x = \mu_y \text{ against } H_1 : \mu_x \neq \mu_y$$

In this paired data, we first obtain the differences:

d_i	–2	–14	–1	0	– 4

Using the differences, we obtain $\bar{d} = -4.2$ and $s_d = 5.67$ Test Statistic:

$$T_c = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{-4.2}{5.67 / \sqrt{5}} = -1.66$$

Using *t*-distribution with 4 d.f. and 1% level and critical values for this two tailed test are – 4 and 4.604. Since the computed value lies in the acceptance region, we cannot reject the null hypoth

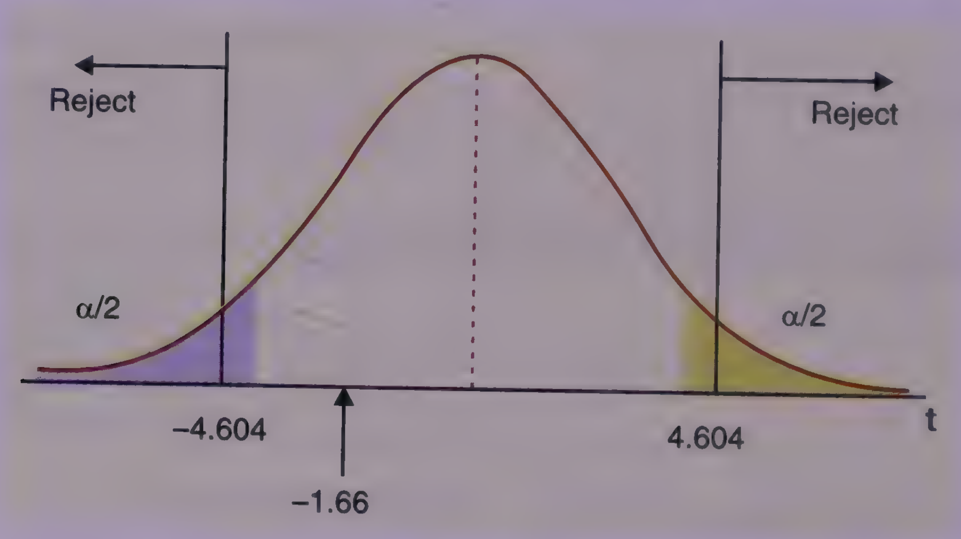


Fig. 10.11

10.11 HYPOTHESIS TESTING FOR DIFFERENCE OF TWO PROPORTION

Consider the following example:

Example 10.8: Random sample of 400 men and 600 women were asked whether they would lik have a flyover near their residence. 200 men and 325 women were in favor of the proposal. Is opinion of men and women the same, in general, on this issue?

In this example we wish to compare two population proportions π_1 and π_2 using sample proporti p_1 and p_2 obtained from two independent samples drawn from each population. Here we may lik test any of the following hypotheses:

(i) $H_0 : \pi_1 = \pi_2$ Against $H_1 : \pi_1 \neq \pi_2$

(ii) $H_0 : \pi_1 = \pi_2$ Against $H_1 : \pi_1 > \pi_2$ or $H_0 : \pi_1 \leq \pi_2$ Against $H_1 : \pi_1 > \pi_2$

(iii) $H_0 : \pi_1 = \pi_2$ Against $H_1 : \pi_1 < \pi_2$ or $H_0 : \pi_1 \geq \pi_2$ Against $H_1 : \pi_1 < \pi_2$

The test statistic is

$$Z_c = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

ere, n_1 and n_2 are the sample sizes and

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Z_c has N(0, 1) distribution provided both sample sizes are large enough (in practice, n_1 and n_2 0).

ution: We wish to test:

$H_0 : \pi_1 = \pi_2$ Against $H_1 : \pi_1 \neq \pi_2$

$p_1 = 200/400 = 0.5, p_2 = 325/600 = 0.5417.$ $\hat{\pi} = \frac{200+325}{400+600} = 0.525$

st statistic:

$$Z_c = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.5 - 0.5417}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = -1.2936.$$

For this two tailed test, at 5% level of significance, the critical values using N(0, 1) distribution, e −1.96 and 1.96. Since the computed value −1.2936 lies in the acceptance region, we cannot ect the null hypothesis. Thus, we cannot say that men and women are different in their opinion.

0.12 HYPOTHESIS TESTING FOR DIFFERENCE OF TWO VARIANCES

onsider the following example:

xample 10.9: You are a financial analyst for a brokerage firm. You collect the following data:

	NYSE	NASDAQ
Number of stocks	21	25
Sample mean	3.27	2.53
Sample std. dev.	1.30	1.16

Is there a difference in the variances (or risks) between the NYSE & NASDAQ?

Suppose that we have two populations with variances σ_1^2 and σ_2^2. We want to examine difference between these two variances. These variances are unknown. So we take a sample f each population and use sample variances to study about population variances. In order to comp two population variances using sample variances, we use Snedecor's *F*-distribution. We wish to t

- $H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$
- $H_0 : \sigma_1^2 \leq \sigma_2^2 \quad H_1 : \sigma_1^2 > \sigma_2^2$
- $H_0 : \sigma_1^2 \geq \sigma_2^2 \quad H_1 : \sigma_1^2 < \sigma_2^2$

The test statistic: $F_c = \dfrac{s_1^2(x)}{s_1^2(y)} = \dfrac{\text{Sample variance of first sample}}{\text{Sample variance of second sample}}$

We write the larger sample variance in the numerator. When both populations are normal, distribution of F_c is Snedecor's *F*-distribution with degree of freedom $(n-1, m-1)$, where n is sample size related to the sample variance used in numerator and m is the sample size of the ot sample. The range of *F* distribution is $(0, \infty)$ and the probability curve of this distribution is n symmetric. Critical values for *F* distribution can be obtained using *F* table (Tables 4(a) and 4(b Appendix) for given d.f. and level of significance.

For two tailed test $(H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2)$:

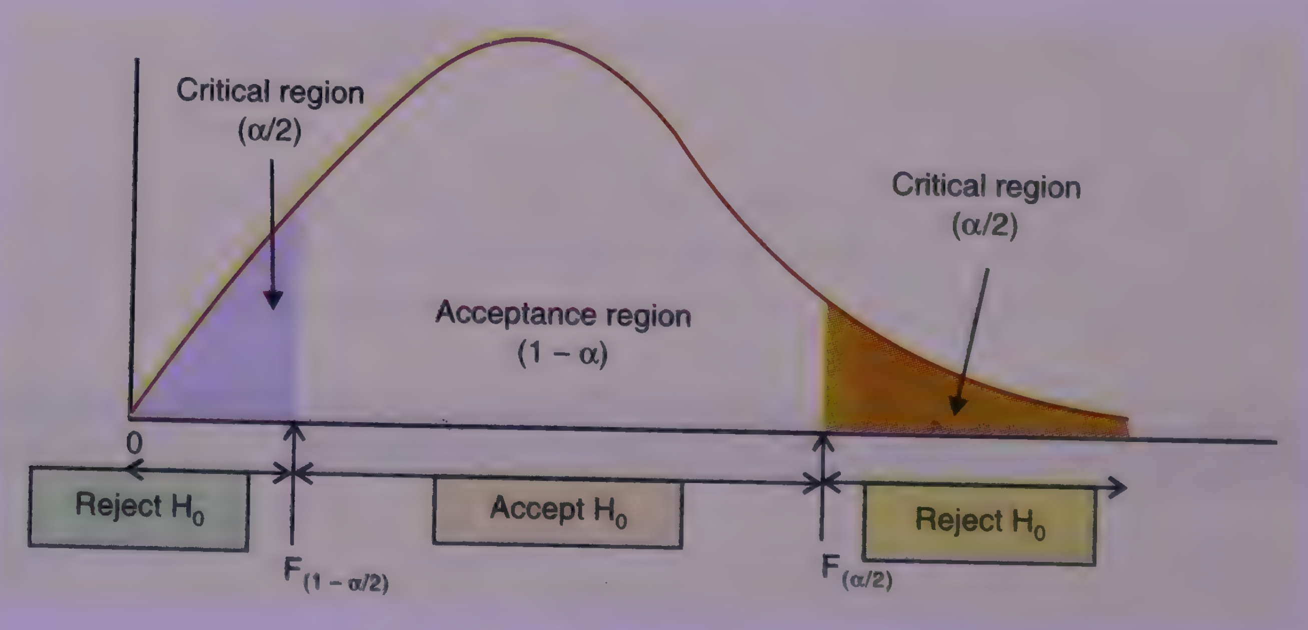


Fig. 10.12

We reject H_0 in the favor of H_a at α ×100% level, if the computed value F_c is smaller tha $F_{(1-\alpha/2)}$ or F_c is greater than $F_{(\alpha\ 2)}$. The notations for critical values are the same as in chi squa distribution.

For right tailed test $(H_0 : \sigma_1^2 = \sigma_2^2$ against $(H_1 : \sigma_1^2 > \sigma_2^2$):

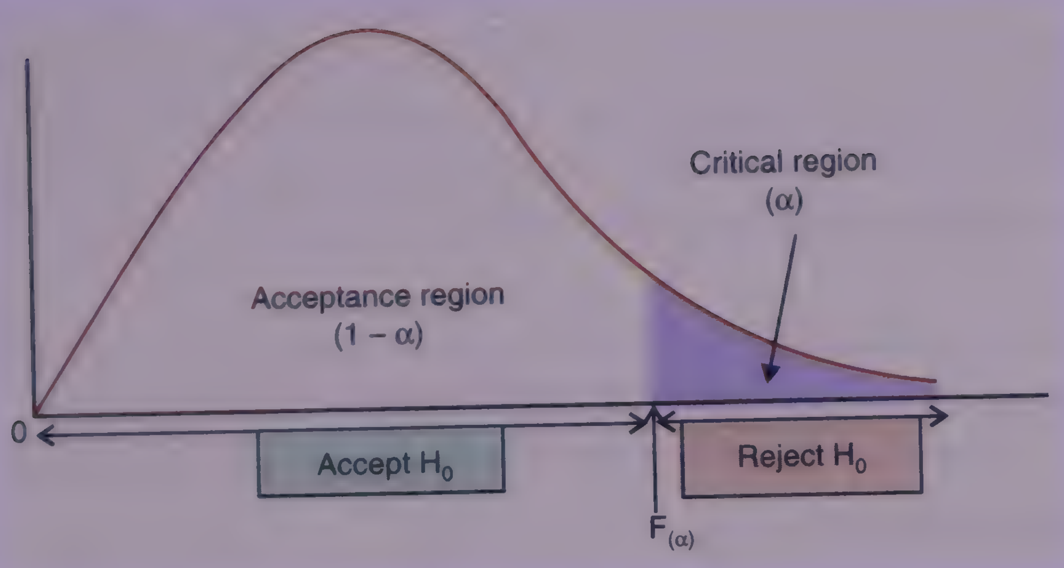

Critical region
(α)
Acceptance region
(1 − α)
0
Accept H₀
Reject H₀
F(α)

Fig. 10.13

We reject H_0 in the favor of H_1 at α × 100% level, if the computed value F_c is greater than $F_{(\alpha)}$.

For left tailed test $(H_0 : \sigma_1^2 = \sigma_2^2$ against $(H_1 : \sigma_1^2 < \sigma_2^2)$.

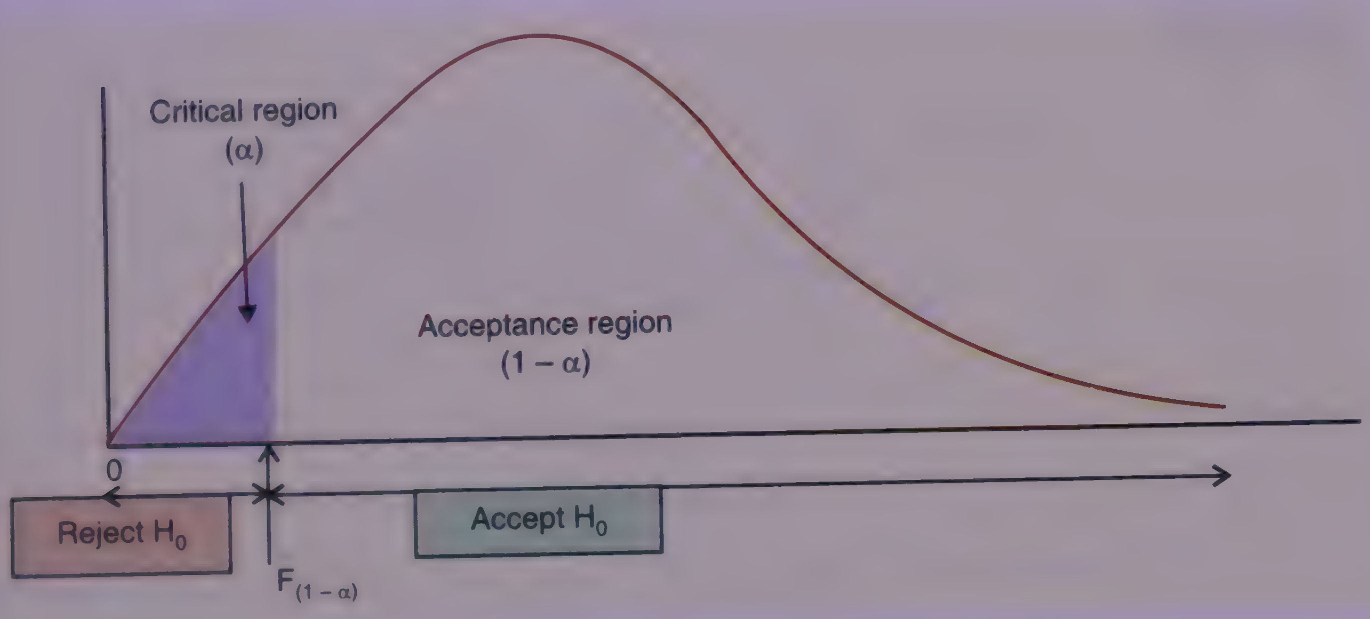

Critical region
(α)
Acceptance region
(1 − α)
0
Reject H₀
Accept H₀
F(1 − α)

Fig. 10.14

We reject H_0 in the favor of H_1 at α × 100% level, if the computed value F_c is smaller than $_{1-\alpha)}$.

Critical values for F distribution can be obtained using F table (Tables 4 (a) and 4 (b) in Appendix) r given d.f. and level of significance. There are two appropriate degrees of freedom—numerator – 1) and denominator ($m - 1$). In the F-table, numerator degrees of freedom determine the column enoted as n_1 or υ_1 usually) and denominator degrees of freedom determine the row (denoted as n_2 υ_2 usually). Also note that

$$F_{(1-\alpha)}(\text{at degree of freedom } \upsilon_1 \text{ and } \upsilon_2) = \frac{1}{F(\alpha)(\text{at degree of freedom } \upsilon_2 \text{ and } \upsilon_1)}$$

This formula is used to obtain critical values sometimes.

Solution of Example 10.9:

We wish to test at 5% level of significance:

$(H_0 : \sigma_1^2 = \sigma_2^2$ Against $H_1 : \sigma_1^2 \neq \sigma_2^2)$

Test statistic:

$$F_c = \frac{s_1^2(x)}{s_1^2(y)} = \frac{\text{Sample variance of first sample}}{\text{Sample variance of second sample}} = \frac{1.30 \times 1.30}{1.60 \times 1.16} = 1.256$$

Critical values at (20, 24) d.f.:

Upper value = $F_U = F_{(0.025)}$ [at d.f. (20, 24)] = 2.33.

Lower value = $F_L = F_{(0.975)}$ [at d.f. (20, 24)] = $1/F_{(0.025)}$ [at d.f. (24, 20)] = 1/2.41 = 0.41.

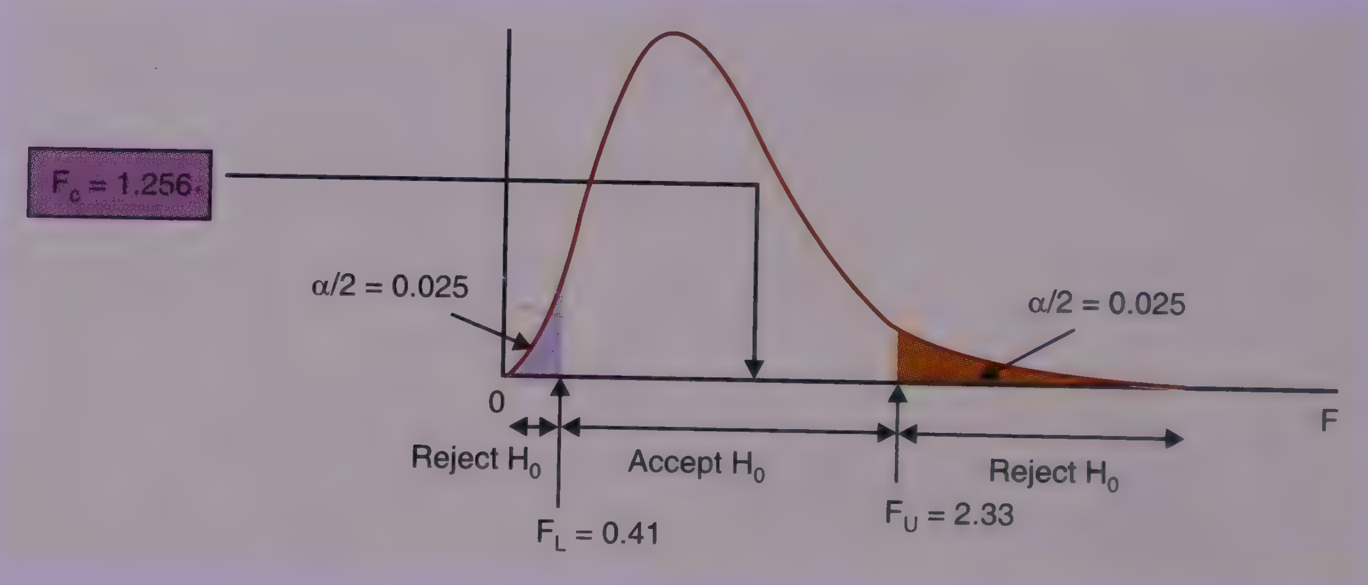


Fig. 10.15

$F_c = 1.256$ is not in the rejection region, so we cannot reject H_0. Therefore, we conclude th there is insufficient evidence of a difference in variances at $\alpha = 0.05$.

10.13 P-VALUE APPROACH

In this approach, we obtain p-value using the computed test statistic, the distribution of test statisti and the type of test (two tailed, right tailed, or left tailed), as below:

Let Z_c be the computed value of test statistic and the distribution of test statistic is N(0, 1), the

- For two-tailed test, p-value = $2P(Z > |Z_c|)$.
- For right-tailed test, p-value = $P(Z > Z_c)$.
- For left-tailed test, p-value = $P(Z < Z_c)$.

where, $Z \sim N(0, 1)$. Above probabilities can be obtained using normal table (Table 1 in Appendix)

Null hypothesis H_0 is rejected in the favor of H_1 at α × 100% level of significance, if ***p-value***

The ***p-value*** is the smallest level of significance at which H_0 would be rejected.

If we wish to see the decision on rejecting or accepting the null hypothesis at different levels of ificance, we need not to obtain different critical values from the table. We can compare various els of significance with the obtained *p*-value. Most of the softwares give the answers in terms of alue.

When the test statistic follows some other distribution, like *t*-distribution, chi-square distribution, -distribution, *p*-value cannot always be obtained using the tables. However, you can obtain alues in such cases using Excel or other softwares.

***mple 10.10*:** A phone industry manager thinks that customer monthly cell phone bills have eased, and now average more than $52 per month. The company wishes to test this claim. pose a sample is taken with the following results Sample size = 64, sample mean = $53.1 per nth. Past company records indicate that the standard deviation is about $10. Test the manager's im at 10% level of significance using *p*-value approach.

ution: Here we wish to test:

$H_0 : \mu \leq 52$ Against $H_1 : \mu > 52$

Population standard deviation is known to be $10. Sample size $n = 64$ which is large enough. We l apply *Z* test for single mean. The test statistic is

$$Z_c = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{53.1 \times 52}{10 / \sqrt{64}} = 0.88$$

The test statistic follows N(0, 1) distribution and it is a right tailed test. Therefore, *p*-value $(Z > 0.88) = 1 - P(Z \leq 0.88) = 1 - [0.5 + P(0 \leq Z \leq 0.88)] = 0.5 - P(0 \leq Z \leq 0.88) = 0.5 - 0.3106$ .1894.

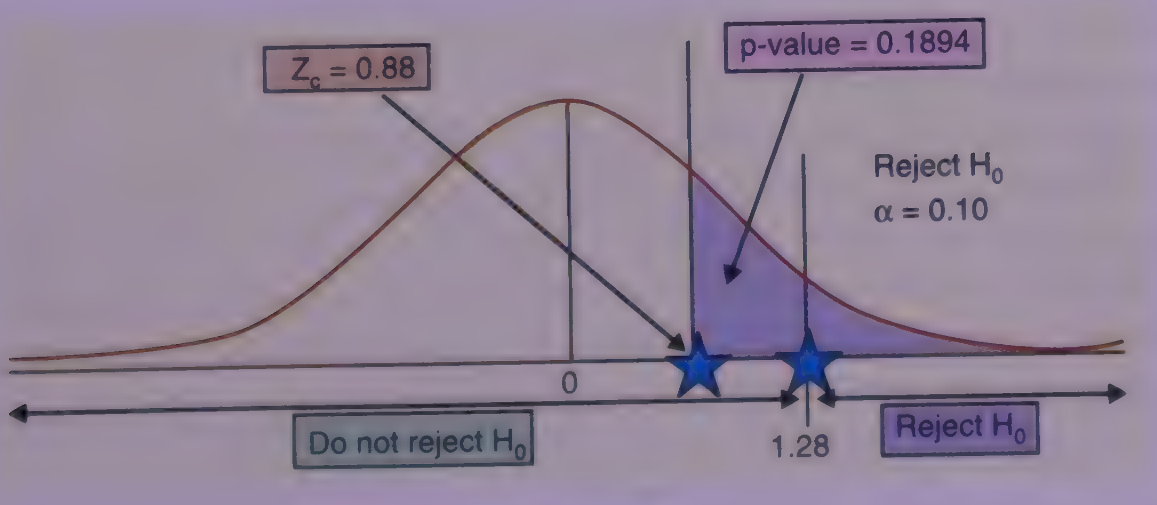


Fig. 10.16

nce *p*-value = 0.1894 > 0.10, we cannot reject null hypothesis at 10% level of significance.

10.14 POWER OF THE TEST

In case we reject H_0, we run a risk of (at most the level of significance) committing an error of I, but if we accept H_0, then we run some risk (the size of which cannot be specified as long as th happens to be vague rather than specific) of committing an error of Type II. The probability of I error is denoted as α (the significance level of the test) and the probability of Type II err referred to as β. Usually the significance level of a test is assigned in advance and once we de it, there is nothing else we can do about α. But what can we say about β? We all know hypothesis test cannot be foolproof; sometimes the test does not reject H_0 when it happens to false one and this way a Type II error is made. But we would certainly like that β (the probabili accepting H_0 when H_0 is not true) to be as small as possible. Alternatively, we would like that 1 (the probability of rejecting H_0 when H_0 is not true) to be as large as possible. If $1 - \beta$ is very m nearer to unity (i.e., nearer to 1.0), we can infer that the test is working quite well, meaning ther that the test is rejecting H_0 when it is not true and if $1 - \beta$ is very much nearer to 0.0, then we i that the test is poorly working, meaning thereby that it is not rejecting H_0 when H_0 is not t Accordingly $1 - \beta$ value is the measure of how well the test is working or what is technic described as the *power of the test*. In case we plot the values of $1 - \beta$ for each possible value of population parameter (say μ, the true population mean) for which the H_0 is not true (alternatively H_1 is true), the resulting curve is known as the power curve associated with the given test. T power curve of a hypothesis test is the curve that shows the conditional probability of rejecting *H* a function of the population parameter and size of the sample.

The function defining this curve is known as the *power function*. In other words, the po function of a test is that function defined for all values of the parameter(s) which yields the probabi that H_0 is rejected and the value of the power function at a specific parameter point is called power of the test at that point. As the population parameter gets closer and closer to hypothesi value of the population parameter, the power of the test (i.e., $1 - \beta$) must get closer and closer to probability of rejecting H_0 when the population parameter is exactly equal to hypothesised valu the parameter. We know that this probability is simply the significance level of the test, and as su the power curve of a test terminates at a point that lies at a height of α (the significance lev directly over the population parameter.

Closely related to the power function, there is another function which is known as the *operat characteristic function* which shows the conditional probability of accepting H_0 for all values population parameter(s) for a given sample size, whether or not the decision happens to be a corr one. If power function is represented as *H* and operating characteristic function as *L*, then we ha $L = 1 - H$. However, one needs only one of these two functions for any decision rule in the cont of testing hypotheses. How to compute the power of a test (i.e., $1 - \beta$) can be explained throu examples.

Example 10.11: A certain chemical process is said to have produced 15 or less pounds of wa material for every 60 lbs. batch with a corresponding standard deviation of 5 lbs. A random sample 100 batches gives an average of 16 lbs of waste per batch. Test at 10 per cent level whether average quantity of waste per batch has increased. Compute the power of the test for $\mu = 16$ lbs ise the level of significance to 20 per cent, then how the power of the test for $\mu = 16$ lbs would ffected?

tion: As we want to test the hypothesis that the average quantity of waste per batch of 60 lbs. or less pounds against the hypothesis that the waste quantity is more than 15 lbs, we can write der:

$$H_0 : \mu \overline{<} 15 \text{ lbs}$$

$$H_1 : \mu > 15 \text{ lbs}$$

As H_1 is one-sided, we shall use the one-tailed test (in the right tail because H_1 is of more than) at 10% level for finding the value of standard deviate (z), corresponding to .4000 area of nal curve which comes to 1.28 as per normal curve area table.* From this we can find the limit of or accepting H_0 as under:

Accept

$$H_0 \text{ if } \quad \overline{X} \overline{<} 15 + 1.28\,(\alpha_p/\sqrt{n})$$

$$\overline{X} \overline{<} 15 + 1.28\left(5/\sqrt{100}\right)$$

$$\overline{X} \overline{<} 15.64$$

0% level of significance otherwise accept H_1.

But the sample average is 16 lbs. which does not come in the acceptance region as above. We, efore, reject H_0 and conclude that average quantity of waste per batch has increased. For finding power of the test, we first calculate β and then subtract it from one. Since β is a conditional bability which depends on the value of μ, we take it as 16 as given in the question. We can now e $\beta = p$ (Accept $H_0 : \mu \overline{<} 15 \,|\, \mu = 16$). Since we have already worked out that H_0 is accepted $\overline{<} 15.64$ (at 10% level of significance), therefore $\beta = p\,(\overline{X} \overline{<} 15.64 \,|\, \mu = 16)$ which can be icted as follows:

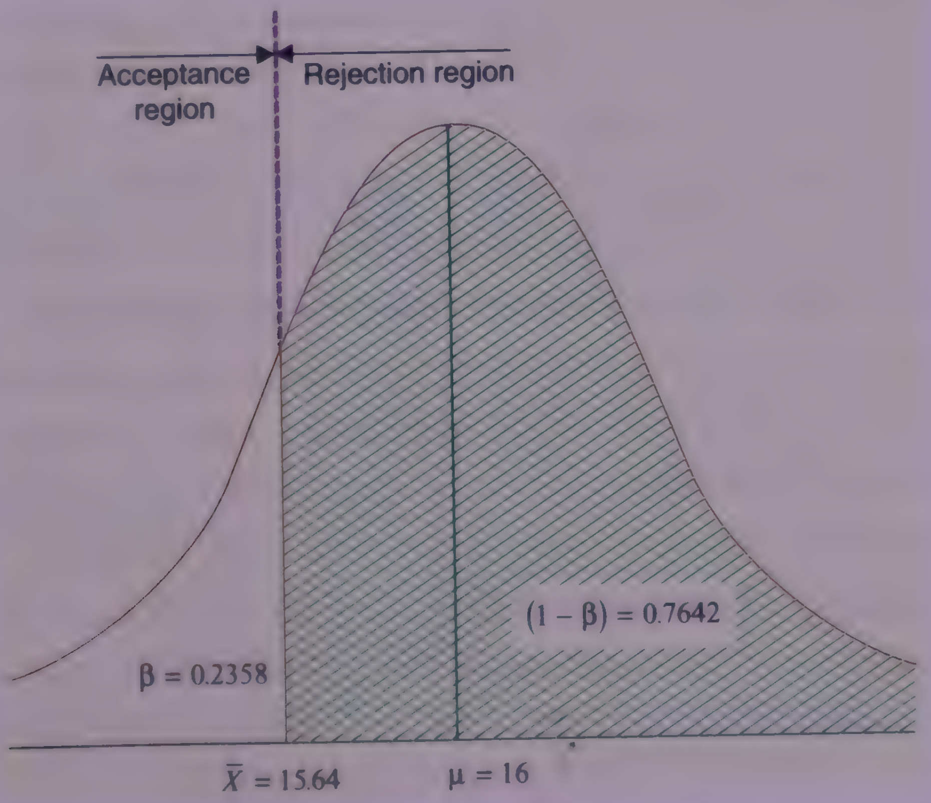


Fig. 10.17

Table No. 1. given in appendix at the end of the book.

We can find out the probability of the area that lies between 15.64 and 16 in the above curve by finding z and then using the area table for the purpose. In the given case $z = (\overline{X} - \mu)/(\sigma/$ $= (15.64 - 16)/(5/\sqrt{100}) = -0.72$ corresponding to which the area is 0.2642. Hence, $\beta = 0.50$ 0.2642 = 0.2358 and the power of the test = $(1 - \beta) = (1 - .2358) = 0.7642$ for $\mu = 16$.

In case the significance level is raised to 20%, then we shall have the following criteria:

Accept H_0 if $\overline{X} \overline{<} 15 + (.84)\left(5/\sqrt{100}\right)$

or $\overline{X} \overline{<} 15.42$, otherwise accept H_1

$\therefore \beta = p\left(\overline{X} \overline{<} 15.42 \mid \mu = 16\right)$

or $\beta = .1230$, , using normal curve area table as explained above.

Hence, $(1-\beta)=(1-0.1230)=0.8770$

10.15 LIMITATIONS OF THE TESTS OF HYPOTHESES

We have described above some important test often used for testing hypotheses on the basi which important decisions may be based. But there are several limitations of the said tests wh should always be borne in mind by a researcher. Important limitations are as follows:

(i) The tests should not be used in a mechanical fashion. It should be kept in view that test is not decision-making itself; the tests are only useful aids for decision-making.

(ii) Test do not explain the reasons as to why does the difference exist, say between the me of the two samples. They simply indicate whether the difference is due to fluctuation sampling or because of other reasons but the tests do not tell us as to which is/are the ot reason(s) causing the difference.

(iii) Results of significance tests are based on probabilities and as such cannot be expres with full certainty. When a test shows that a difference is statistically significant, the simply suggests that the difference is probably not due to chance.

(iv) Statistical inferences based on the significance tests cannot be said to be entirely corr evidences concerning the truth of the hypotheses. This is specially so in case of sm samples where the probability of drawing erring inferences happens to be generally high For greater reliability, the size of samples be sufficiently enlarged.

All these limitations suggest that in problems of statistical significance, the inference techniqu (or the tests) must be combined with adequate knowledge of the subject-matter along with the abil of good judgement.

e Solved Examples

mple 10.12: A sample of 400 male students is found to have a mean height 67.47 inches. Can reasonably regarded as a sample from a large population with mean height 67.39 inches and dard deviation 1.30 inches? Test at 5% level of significance.

tion: Taking the null hypothesis that the mean height of the population is equal to 67.39 inches, an write:

$$H_0 : \mu = 67.39''$$

$$H_1 : \mu \neq 67.39''$$

the given information as $\overline{X} = 67.47''$, $\sigma = 1.30''$, $n = 400$. Assuming the population to be nal, we can work out the test statistic z as under:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{67.47 - 67.39}{1.30/\sqrt{400}} = \frac{0.08}{0.065} = 1.231$$

As H_1 is two-sided in the given question, we shall be applying a two-tailed test for determining rejection regions at 5% level of significance which comes to as under, using normal curve area e:

$$R : |z| > 1.96$$

The observed value of z is 1.231 which is in the acceptance region since $R : |z| > 1.96$ and thus s accepted. We may conclude that the given sample (with mean height = 67.47″) can be regarded ave been taken from a population with mean height 67.39″ and standard deviation 1.30″ at 5% el of significance.

mple 10.13: Suppose we are interested in a population of 20 industrial units of the same size, all which are experiencing excessive labour turnover problems. The past records show that the mean he distribution of annual turnover is 320 employees, with a standard deviation of 75 employees. A nple of 5 of these industrial units is taken at random which gives a mean of annual turnover as 300 ployees. Is the sample mean consistent with the population mean? Test at 5% level.

ution: Taking the null hypothesis that the population mean is 320 employees, we can write:

$$H_0 : \mu = 320$$

$$H_1 : \mu \neq 320$$

l the given information as under:

$$\overline{X} = 300 \text{ employees}, \ \sigma = 75 \text{ employees}$$
$$n = 5;\ N = 20$$

Assuming the population to be normal, we can work out the test statistic z as under:

$$z^* = \frac{\overline{X} - \mu}{\sigma/\sqrt{n} \times \sqrt{(N-n)/(N-1)}}$$

---

[*] Being a case of finite population.

$$= \frac{300 - 320}{75/\sqrt{5} \times \sqrt{(20-5)/(20-1)}} = -\frac{20}{(33.54)\,(.888)}$$

$$= -0.67$$

As H_1 is two-sided in the given question, we shall apply a two-tailed test for determining rejection regions at 5% level of significance which comes to as under, using normal curve area ta

$$R : |z| > 1.96$$

The observed value of z is -0.67 which is in the acceptance region since $R : |z| > 1.96$ and tl H_0 is accepted and we may conclude that the sample mean is consistent with population mean the population mean 320 is supported by sample results.

Example 10.14: The mean of a certain production process is known to be 50 with a stanc deviation of 2.5. The production manager may welcome any change is mean value towards hig side but would like to safeguard against decreasing values of mean. He takes a sample of 12 ite that gives a mean value of 48.5. What inference should the manager take for the production proc on the basis of sample results? Use 5 per cent level of significance for the purpose.

Solution: Taking the mean value of the population to be 50, we may write:

$$H_o : \mu = 50$$

$H_1 : \mu < 50$ (Since the manager wants to safeguard against decreasing values of mean.

and the given information as $\overline{X} = 48.5$, $\sigma = 2.5$ and $n = 12$. Assuming the population to be norm we can work out the test statistic z as under:

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{48.5 - 50}{2.5/\sqrt{12}} = -\frac{1.5}{2.5/3.464} = -2.0784$$

As H_1 is one-sided in the given question, we shall determine the rejection region applying o tailed test (in the left tail because H_1 is of less than type) at 5 per cent level of significance anc comes to as under, using normal curve area table:

$$R : z < -1.645$$

The observed value of z is -2.0784 which is in the rejection region and thus, H_0 is rejected a per cent level of significance. We can conclude that the production process is showing mean whi is significantly less than the population mean and this calls for some corrective action concerning t said process.

Example 10.15: The specimen of copper wires drawn from a large lot have the following breaki strength (in kg. weight):

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Test (using Student's t-statistic) whether the mean breaking strength of the lot may be taken be 578 kg. weight (Test at 5 per cent level of significance). Verify the inference so drawn by usi Sandler's A-statistic as well.

tion: Taking the null hypothesis that the population mean is equal to hypothesised mean of kg, we can write:

$$H_0 : \mu = 578 \text{ kg.}$$
$$H_1 : \mu \neq 578 \text{ kg.}$$

As the sample size is small (since $n = 10$) and the population standard deviation is not known, we l use t-test assuming normal population and shall work out the test statistic t as under:

$$t = \frac{\bar{X} - \mu_0}{s_1 / \sqrt{n}}$$

To find $\bar{X}$ and s_1 we make the following computations:

Table 10.4

S. No.	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	578	6	36
2	572	0	0
3	570	– 2	4
4	568	– 4	16
5	572	0	0
6	578	6	36
7	570	– 2	4
8	572	0	0
9	596	24	576
10	544	– 28	784
$n = 10$	$\Sigma X_i = 5720$		$\Sigma(X_i - \bar{X})^2 = 1456$

$$\therefore \quad \bar{X} = \frac{\Sigma X_i}{n} = \frac{5720}{10} = 572 \text{ kg}$$

d

$$s_1 = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{1456}{10-1}} = 12.72 \text{ kg.}$$

Hence, $$t = \frac{572 - 578}{12.72/\sqrt{10}} = -1.488$$

Degree of freedom = $(n - 1) = (10 - 1) = 9$

As H_1 is two-sided, we shall determine the rejection region applying two-tailed test at 5 per cent vel of significance, and it comes to as under, using table of t-distribution for 9 d.f.:

$$R: |t| > 2.262$$

As the observed value of t (i.e., – 1.488) is in the acceptance region, we accept H_0 at 5 per level and conclude that the mean breaking strength of copper wires lot may be taken as 578 weight.

The same inference can be drawn using Sandler's A-statistic as shown below:

Table 10.5: Computations for A-Statistic

S. No.	X_i	Hypothesised mean $\mu_0 = 578$ kg.	$D_i = (X_i - \mu_0)$	D_i^2
1	578	578	0	0
2	572	578	–6	36
3	570	578	–8	64
4	568	578	–10	100
5	572	578	–6	36
6	578	578	0	0
7	570	578	–8	64
8	572	578	–6	36
9	596	578	18	324
10	544	578	–34	1156
$n = 10$			$\Sigma D_i = -60$	$\Sigma D_i^2 = 181$

$$\therefore \qquad A = \Sigma D_i^2 / (\Sigma D_i)^2 = 1816/(-60)^2 = 0.5044$$

Null hypothesis $\quad H_0 : \mu = 578$ kg

Alternate hypothesis $H_1 : \mu \neq 578$ kg

As H_1 is two-sided, the critical value of A-statistic from the A-statistic table (Table No. 11 giv in appendix at the end of the book) for $(n - 1)$ i.e., $10 - 1 = 9$ d.f. at 5% level is 0.276. Compu value of A (0.5044), being greater than 0.276 shows that A-statistic is insignificant in the given ca and accordingly we accept H_0 and conclude that the mean breaking strength of copper wire' maybe taken as 578 kg weight. Thus, the inference on the basis of t-statistic stands verified A-statistic.

Example 10.16: Raju Restaurant near the railway station at Falna has been having average sales 500 tea cups per day. Because of the development of bus stand nearby, it expects to increase sales. During the first 12 days after the start of the bus stand, the daily sales were as under:

550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526

On the basis of this sample information, can one conclude that Raju Restaurant's sales ha increased? Use 5 per cent level of significance.

tion: Taking the null hypothesis that sales average 500 tea cups per day and they have not eased unless proved, we can write:

$H_0 : \mu = 500$ cups per day

$H_1 : \mu > 500$ (as we want to conclude that sales have increased).

As the sample size is small and the population standard deviation is not known, we shall use *t*-test ming normal population and shall work out the test statistic *t* as:

$$t = \frac{\bar{X} - \mu_0}{s_1/\sqrt{n}}$$

(To find $\bar{X}$ and s_1 we make the following computations):

Table 10.6

S. No.	X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	550	2	4
2	570	22	484
3	490	–58	3364
4	615	67	4489
5	505	–43	1849
6	580	32	1024
7	570	22	484
8	460	–88	7744
9	600	52	2704
10	580	32	1024
11	530	–18	324
12	526	–22	484
$n = 10$	$\Sigma X_i = 6576$		$\Sigma(X_i - \bar{X})^2 = 23978$

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{6576}{12} = 548$$

nd

$$s_1 = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

ence,

$$t = \frac{548 - 500}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

Degree of freedom $= n - 1 = 12 - 1 = 11$

As H_1 is one-sided, we shall determine the rejection region applying one-tailed test (in the r tail because H_1 is of more than type) at 5 per cent level of significance and it comes to as under, u table of t-distribution for 11 degrees of freedom:

$$R : t > 1.796$$

The observed value of t is 3.558 which is in the rejection region and thus H_0 is rejected at 5 cent level of significance and we can conclude that the sample data indicate that Raju restaura sales have increased.

Example 10.17: The mean produce of wheat of a sample of 100 fields in 200 lbs. per acre wi standard deviation of 10 lbs. Another samples of 150 fields gives the mean of 220 lbs. with a stand deviation of 12 lbs. Can the two samples be considered to have been taken from the same populat whose standard deviation is 11 lbs? Use 5 per cent level of significance.

Solution: Taking the null hypothesis that the means of two populations do not differ, we can wri

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

and the given information as $n_1 = 100; n_2 = 150;$

$$\bar{X}_1 = 200 \text{ lbs} \quad \bar{X}_2 = 220 \text{ lbs}$$

$$s_1 = 10 \text{ lbs} \quad s_2 = 12 \text{ lbs}$$

and

$$\sigma = 11 \text{ lbs}$$

Assuming the population to be normal, we can work out the test statistic z as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{200 - 220}{\sqrt{11^2\left(\frac{1}{100} + \frac{1}{150}\right)}}$$

$$= -\frac{20}{1.42} = -14.08$$

As H_1 is two-sided, we shall apply a two-tailed test for determining the rejection regions at 5 pe cent level of significance which come to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is –14.08 which falls in the rejection region and thus we reject H_0 an conclude that the two samples cannot be considered to have been taken at 5 per cent level c significance from the same population whose standard deviation is 11 lbs. This means that the differenc between means of two samples is statistically significant and not due to sampling fluctuations.

Example 10.18: A simple random sampling survey in respect of monthly earnings of semi-skille workers in two cities gives the following statistical information:

Table 10.7

City	Mean monthly earnings (₹)	Standard deviation of sample data of monthly earnings (₹)	Size of sample
A	695	40	200
B	710	60	175

Test the hypothesis at 5 per cent level that there is no difference between monthly earnings of kers in the two cities.

ution: Taking the null hypothesis that there is no difference in earnings of workers in the two es, we can write:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

d the given information as

Sample 1 (City A)	*Sample 2 (City B)*
$\bar{X}_1 =$ ₹ 695	$\bar{X}_2 =$ ₹ 710
$s_1 =$ ₹ 40	$s_2 =$ ₹ 60
$n_1 = 200$	$n_2 = 175$

s the sample size is large, we shall use *z*-test for difference in means assuming the populations to be rmal and shall work out the test statistic *z* as under:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Since the population variances are not known, we have used the sample variances, considering e sample variances as the estimates of population variances.)

Hence $$z = \frac{695 - 710}{\sqrt{\frac{(40)^2}{200} + \frac{(60)^2}{175}}} = -\frac{15}{\sqrt{8 + 20.57}} = -2.809$$

As H_1 is two-sided, we shall apply a two-tailed test for determining the rejection regions at 5 per ent level of significance which come to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is -2.809 which falls in the rejection region and thus we reject H_0 at per cent level and conclude that earning of workers in the two cities differ significantly.

Example 10.19: Sample of sales in similar shops in two towns are taken for a new product with following results:

Town	*Mean sales*	*Variance*	*Size of sample*
A	57	5.3	5
B	61	4.8	7

Is there any evidence of difference in sales in the two towns? Use 5 per cent level of significan for testing this difference between the means of two samples.

Solution: Taking the null hypothesis that the means of two populations do not differ we can write

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

and the given information as follows:

Table 10.8

Sample from town A as sample one	$\bar{X}_1 = 57$	$s_1^2 = 53$	$n_1 = 5$
Sample from town B as sample two	$\bar{X}_2 = 61$	$s_2^2 = 48$	$n_2 = 7$

Since in the given question variances of the population are not known and the size of samples i small, we shall use *t*-test for difference in means, assuming the populations to be normal and ca work out the test statistic *t* as under:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with d.f. = $(n_1 + n_2 - 2)$

$$= \frac{57 - 61}{\sqrt{\frac{4(5.3) + 6(4.8)}{5 + 7 - 2}} \times \sqrt{\frac{1}{5} + \frac{1}{7}}} = -3.053$$

Degrees of freedom = $(n_1 + n_2 - 2) = 5 + 7 - 2 = 10$

As H_1 is two-sided, we shall apply a two-tailed test for determining the rejection regions at 5 per cent level which come to as under, using table of *t*-distribution for 10 degrees of freedom:

$$R : |t| > 2.228$$

The observed value of *t* is –3.053 which falls in the rejection region and thus, we reject H_0 and conclude that the difference in sales in the two towns is significant at 5 per cent level.

ample 10.20: A group of seven-week old chickens reared on a high protein diet weigh 12, 15, 11, , 14, 14, and 16 ounces; a second group of five chickens, similarly treated except that they receive ow protein diet, weigh 8, 10, 14, 10 and 13 ounces. Test at 5 per cent level whether there is gnificant evidence that additional protein has increased the weight of the chickens. Use assumed ean (or A_1) = 10 for the sample of 7 and assumed mean (or A_2) = 8 for the sample of 5 chickens in ur calculations.

lution: Taking the null hypothesis that additional protein has not increased the weight of the chickens e can write:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$ (as we want to conclude that additional protein has increased the weight of chickens)

Since in the given question variances of the populations are not known and the size of samples is nall, we shall use *t*-test for difference in means, assuming the populations to be normal and thus ork out the test statistic *t* as under:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

vith d.f. = $(n_1 + n_2 - 2)$

From the sample data we work out $\bar{X}_1$, $\bar{X}_2$, s_1^2 and s_2^2 (taking high protein diet sample as sample ne and low protein diet sample as sample two) as shown below:

Table 10.9

Sample one				Sample two			
S.No.	X_{1i}	$X_{1i} - A_1$ $(A_1 = 10)$	$(X_{1i} - A_1)^2$	S.No.	X_{2i}	$X_{2i} - A_2$ $(A_2 = 8)$	$(X_{2i} - A_2)^2$
1	12	2	4	1	8	0	0
2	15	5	25	2	10	2	4
3	11	1	1	3	14	6	36
4	16	6	36	4	10	2	4
5	14	4	16	5	13	5	25
6	14	4	16				
7	16	6	36				
$n_1 = 7$;	$\Sigma(X_{1i} - A_1) = 28$;		$\Sigma(X_{1i} - A_1)^2 = 134$	$n_2 = 5$;	$\Sigma(X_{2i} - A_2) = 15$;		$\Sigma(X_{2i} - A_2)^2 = 69$

$$\therefore \qquad \bar{X}_1 = A_1 + \frac{\Sigma(X_{1i} - A_1)}{n_1} = 10 + \frac{28}{7} = 14 \text{ ounces}$$

$$\bar{X}_2 = A_2 + \frac{\Sigma(X_{2i} - A_2)}{n_2} = 8 + \frac{15}{5} = 11 \text{ ounces}$$

$$s_1^2 = \frac{\Sigma(X_{1i} - A_1)^2 - \left[\Sigma(X_{1i} - A_1)\right]^2 / n_1}{(n_1 - 1)}$$

$$= \frac{134 - (28)^2/7}{7 - 1} = 3.667 \text{ ounces}$$

$$s_2^2 = \frac{\Sigma(X_{2i} - A_2)^2 - \left[\Sigma(X_{2i} - A_2)\right]^2 / n_2}{(n_2 - 1)}$$

$$= \frac{69 - (15)^2/5}{5 - 1} = 6 \text{ ounces}$$

Hence,

$$t = \frac{14 - 11}{\sqrt{\dfrac{(7-1)(3.667) + (5-1)(6)}{7 + 5 - 2}} \times \sqrt{\dfrac{1}{7} + \dfrac{1}{5}}}$$

$$= \frac{3}{\sqrt{4.6} \times \sqrt{.345}} = \frac{3}{1.26} = 2.381$$

Degrees of freedom = $(n_1 + n_2 - 2) = 10$

As H_1 is one-sided, we shall apply a one-tailed test (in the right tail because H_1 is of more tha type) for determining the rejection region at 5 per cent level which comes to as under, using table c t-distribution for 10 degrees of freedom:

$$R : t > 1.812$$

The observed value of t is 2.381 which falls in the rejection region and thus, we reject H_0 an conclude that additional protein has increased the weight of chickens, at 5 per cent level of significanc

Example 10.21: Memory capacity of 9 students was tested before and after training. State at 5 pe cent level of significance whether the training was effective from the following scores:

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

Use paired t-test as well as A-test for your answer.

lution: Take the score before training as X and the score after training as Y and then taking the null pothesis that the mean of difference is zero, we can write:

$H_0 : \mu_1 = \mu_2$ which is equivalent to test $H_0 : \bar{D} = 0$

$H_1 : \mu_1 < \mu_2$ (as we want to conclude that training has been effective)

As we are having matched pairs, we use paired t-test and work out the test statistic t as under:

$$t = \frac{\bar{D} - 0}{\sigma_{diff.}/\sqrt{n}}$$

find the value of t, we shall first have to work out the mean and standard deviation of differences shown below:

Table 10.10

Student	Score before training X_i	Score after training Y_i	Difference $(D_i = X_i - Y_i)$	Difference Squared D_i^2
1	10	12	−2	4
2	15	17	−2	4
3	9	8	1	1
4	3	5	−2	4
5	7	6	1	1
6	12	11	1	1
7	16	18	−2	4
8	17	20	−3	9
9	4	3	1	1
$n = 9$			$\Sigma D_i = -7$	$\Sigma D_i^2 = 29$

$\therefore$ Mean of Differences or $\bar{D} = \dfrac{\Sigma D_i}{n} = \dfrac{-7}{9} = -0.778$

nd Standard deviation of differences or

$$\sigma_{diff} = \sqrt{\frac{\sum D_i^2 - (\bar{D})^2 \cdot n}{n-1}}$$

$$= \sqrt{\frac{29 - (-.778)^2 \times 9}{9-1}}$$

$$= \sqrt{2.944} = 1.715$$

Hence, $$t = \frac{-0.778 - 0}{1.715/\sqrt{9}} = \frac{-.778}{0.572} = -1.361$$

Degrees of freedom $= n - 1 = 9 - 1 = 8$.

As H_1 is one-sided, we shall apply a one-tailed test (in the left tail because H_1 is of less than ty
for determining the rejection region at 5 per cent level which comes to as under, using the table
t-distribution for 8 degrees of freedom:

$$R : t < -1.860$$

The observed value of t is -1.361 which is in the acceptance region and thus, we accept H_0 a
conclude that the difference in score before and after training is insignificant i.e., it is only due
sampling fluctuations. Hence we can infer that the training was not effective.

Solution using A-test: Using A-test, we workout the test statistic for the given problem thus

$$A = \frac{\Sigma D_i^2}{(\Sigma D_i)^2} = \frac{29}{(-7)^2} = 0.592$$

Since H_1 in the given problem is one-sided, we shall apply one-tailed test. Accordingly, at 5
level of significance the table value of A-statistic for $(n - 1)$ or $(9 - 1) = 8$ d.f. in the given case
0.368 (as per table of A-statistic given in appendix). The computed value of A i.e., 0.592 is high
than this table value and as such A-statistic is insignificant and accordingly H_0 should be accepted.
other words, we should conclude that the training was not effective. (This inference is just the sam
as drawn earlier using paired t-test.)

Example 10.22: The sales data of an item in six shops before and after a special promotion
campaign are:

Shops	*A*	B	C	D	E	F
Before the promotional campaign	53	28	31	48	50	42
After the campaign	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5 per cent level of significance. Use paire
t-test as well as A-test.

Solution: Let the sales before campaign be represented as X and the sales after campaign as Y an
then taking the null hypothesis that campaign does not bring any improvement in sales, we can write

$H_0 : \mu_1 = \mu_2$ which is equivalent to test $H_0 : \overline{D} = 0$

$H_1 : \mu_1 < \mu_2$ (as we want to conclude that campaign has been a success).

Because of the matched pairs we use paired t-test and work out the test statistic 't' as under:

$$t = \frac{\overline{D} - 0}{\sigma_{diff} / \sqrt{n}}$$

To find the value of t, we first work out the mean and standard deviation of differences as under:

Table 10.11

Shops	Sales before campaign X_i	Sales after campaign Y_i	Difference $(D_i = X_i - Y_i)$	Difference squared D_i^2
A	53	58	−5	25
B	28	29	−1	1
C	31	30	1	1
D	48	55	−7	49
E	50	56	−6	36
F	42	45	−3	9
$n = 6$			$\Sigma D_i = -21$	$\Sigma D_i^2 = 121$

$$\therefore \qquad \overline{D} = \frac{\Sigma D_i}{n} = -\frac{21}{6} = -3.5$$

$$\sigma_{diff} = \sqrt{\frac{\sum D_i^2 - (\overline{D})^2 \cdot n}{n-1}} = \sqrt{\frac{121 - (-3.5)^2 \times 6}{6-1}} = 3.08$$

Hence,
$$t = \frac{-3.5 - 0}{3.08/\sqrt{6}} = \frac{-3.5}{1.257} = -2.784$$

Degrees of freedom = $(n - 1) = 6 - 1 = 5$

As H_1 is one-sided, we shall apply a one-tailed test (in the left tail because H_1 is of less than type) or determining the rejection region at 5 per cent level of significance which come to as under, using able of t-distribution for 5 degrees of freedom:

$$R : t < -2.015$$

The observed value of t is −2.784 which falls in the rejection region and thus, we reject H_0 at 5 er cent level and conclude that sales promotional campaign has been a success.

Jsing A-test: Using A-test, we work out the test statistic for the given problem as under:

$$A = \frac{\Sigma D_i^2}{(\Sigma D_i)^2} = \frac{121}{(-21)^2} = 0.2744$$

Since H_1 in the given problem is one-sided, we shall apply one-tailed test. Accordingly, at 5% evel of significance the table value of A-statistic for $(n - 1)$ or $(6 - 1) = 5$ d.f. in the given case is .372 (as per table of A-statistic given in appendix). The computed value of A, being 0.2744, is less han this table value and as such A-statistic is significant. This means we should reject H_0 (alternately ve should accept H_1) and should infer that the sales promotional campaign has been a success.

Example 10.23: A sample survey indicates that out of 3232 births, 1705 were boys and the r were girls. Do these figures confirm the hypothesis that the sex ratio is 50 : 50? Test at 5 per c level of significance.

Solution: Starting from the null hypothesis that the sex ratio is 50 : 50 we may write:

$$H_o : \pi = \frac{1}{2}$$

$$H_1 : \pi \neq \frac{1}{2}$$

Hence the probability of boy birth or $\pi = \frac{1}{2}$ and the probability of girl birth is also $\frac{1}{2}$.

Considering boy birth as success and the girl birth as failure, we can write as under:

The proportion of success or $\pi = \frac{1}{2}$

The proportion of failure or $1 - \pi = \frac{1}{2}$

and $n = 3232$ (given).

The standard error of proportion of success.

$$= \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{3232}} = 0.0088$$

Observed sample proportion of success, or

$$\hat{\pi} = p = 1705 / 3232 = 0.5275$$

and the test statistic

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1-\pi_0)}{n}}} = \frac{0.5275 - 0.5000}{0.0088} = 3.125$$

As H_1 is two-sided in the given question, we shall be applying the two-tailed test for determining the rejection regions at 5 per cent level which come to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is 3.125 which comes in the rejection region since $R : |z| > 1.96$ and thus, H_0 is rejected in favour of H_1. Accordingly, we conclude that the given figures do not conform the hypothesis of sex ratio being 50 : 50.

ample 10.24: The null hypothesis is that 20 per cent of the passengers go in first class, but anagement recognizes the possibility that this percentage could be more or less. A random sample 400 passengers includes 70 passengers holding first class tickets. Can the null hypothesis be jected at 10 per cent level of significance?

lution: The null hypothesis is

$H_0 : \pi = 0.20$

nd $H_1 : \pi \neq 0.20$

Observed sample proportion $(p) = 70/400 = 0.175$

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.175 - 0.20}{\sqrt{\frac{0.20 \times 0.80}{400}}} = -1.25$$

nd the test statistic

As H_1 is two-sided we shall determine the rejection regions applying two-tailed test at 10 per ent level which come to as under, using normal curve area table:

$$R : |z| > 1.645$$

The observed value of z is -1.25 which is in the acceptance region and as such H_0 is accepted. hus the null hypothesis cannot be rejected at 10 per cent level of significance.

xample 10.25: A certain process produces 10 per cent defective articles. A supplier of new raw aterial claims that the use of his material would reduce the proportion of defectives. A random ample of 400 units using this new material was taken out of which 34 were defective units. Can the upplier's claim be accepted? Test at 1 per cent level of significance.

olution: The null hypothesis can be written as $H_0 : \pi = 10\%$ or 0.10 and the alternative hypothesis $H_1 : \pi < 0.10$ (because the supplier claims that new material will reduce proportion of defectives).

Observed sample proportion $p = 34/400 = 0.085$ and test statistic

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.085 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{400}}} = \frac{-0.015}{0.015} = -1.00$$

As H_1 is one-sided, we shall determine the rejection region applying one-tailed test (in the left tail because H_1 is of less than type) at 1% level of significance and it comes to as under, using normal curve area table:

$$R : z < -2.32$$

As the computed value of z does not fall in the rejection region, H_0 is accepted at 1% level of significance and we can conclude that on the basis of sample information, the supplier's claim cannot be accepted at 1% level.

Example 10.26: A drug research experimental unit is testing two drugs newly developed to redu blood pressure levels. The drugs are administered to two different sets of animals. In group one, 3 of 600 animals tested respond to drug one and in group two, 260 of 500 animals tested respond drug two. The research unit wants to test whether there is a difference between the efficacy of t said two drugs at 5 per cent level of significance. How will you deal with this problem?

Solution: We take the null hypothesis that there is no difference between the two drugs i.e., $H_0 : \pi_1 =$

The alternative hypothesis can be taken as that there is a difference between the drugs i. $H_1 : \pi_1 = \pi_2$ and the given information can be stated as:

$$p_1 = 350/600 = 0.583$$

$$q_1 = 1 - p_1 = 0.417$$

$$n_1 = 600$$

$$p_2 = 260/500 = 0.520$$

$$q_2 = 1 - p_2 = 0.480$$

$$n_2 = 500$$

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = 0.555$$

We can work out the test statistic z thus:

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= 2.094$$

As H_1 is two-sided, we shall determine the rejection regions applying two-tailed test at 5% leve which comes as under using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is 2.093 which is in the rejection region and thus, H_0 is rejected in favour o H_1 and as such we conclude that the difference between the efficacy of the two drugs is significant

Example 10.27: At a certain date in a large city 400 out of a random sample of 500 men were founc to be smokers. After the tax on tobacco had been heavily increased, another random sample of 60 men in the same city included 400 smokers. Was the observed decrease in the proportion of smoker significant? Test at 5 per cent level of significance.

lution: We start with the null hypothesis that the proportion of smokers even after the heavy tax tobacco remains unchanged i.e. $H_o : \pi_1 = \pi_2$ and the alternative hypothesis that proportion of okers after tax has decreased i.e.,

$$H_1 : \pi_1 > \pi_2$$

On the presumption that the given populations are similar as regards the given attribute, we work t the best estimate of proportion of smokers (p_0) in the population as under, using the given information:

$$\hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{500\left(\frac{400}{500}\right) + 600\left(\frac{400}{600}\right)}{500 + 600} = \frac{800}{1100} = \frac{8}{11} = .7273$$

The test statistic z can be worked out as under:

$$z = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.133}{0.027} = 4.926$$

As the H_1 is one-sided we shall determine the rejection region applying one-tailed test (in the ght tail because H_1 is of greater than type) at 5 per cent level and the same works out to as under, sing normal curve area table:

$$R : z > 1.645$$

The observed value of z is 4.926 which is in the rejection region and so we reject H_0 in favour of H_1 and conclude that the proportion of smokers after tax has decreased significantly.

Testing the difference between proportion based on the sample and the proportion given or the whole population: In such a situation we work out the standard error of difference between roportion of persons possessing an attribute in a sample and the proportion given for the population s under:

Standard error of difference between sample proportion and population proportion or

$$.E._{diff\ \hat{p}-p} = \sqrt{\pi(1-\pi)\frac{N-n}{nN}}$$

here, π = Population proportion

n = Number of items in the sample

N = Number of items in population

nd the test statistic z can be worked out as under:

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)\frac{N-n}{nN}}}$$

All other steps remain the same as explained above in the context of testing of proportions. take an example to illustrate the same.

Example 10.28: There are 100 students in a university college and in the whole university, inclusi of this college, the number of students is 2000. In a random sample study 20 were found smokers the college and the proportion of smokers in the university is 0.05. Is there a significant differen between the proportion of smokers in the college and university? Test at 5 per cent level.

Solution: Let $H_0 : \pi = 100/2000 = 0.05$

and $H_1 : \pi \neq 0.05$

and on the basis of the given information, the test statistic z can be worked out as under:

$$z = \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)\frac{N-n}{Nn}}} = \frac{\frac{20}{100} - 0.05}{\sqrt{(0.05)(0.95)\frac{2000-100}{(100)(2000)}}}$$

$$= \frac{0.150}{0.021} = 7.143$$

As the H_1 is two-sided, we shall determine the rejection regions applying two-tailed test at 5 p cent level and the same works out to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is 7.143 which is in the rejection region and as such we reject H_0 an conclude that there is a significant difference between the proportion of smokers in the college an university.

Example 10.29: Two random samples drawn from two normal populations are:

Sample 1	20	16	26	27	23	22	18	24	25	19		
Sample 2	27	33	42	35	32	34	38	28	41	43	30	37

Test using variance ratio at 5 per cent and 1 per cent level of significance whether the tw populations have the same variances.

ution: We take the null hypothesis that the two populations from where the samples have been wn have the same variances i.e., $H_0 : \sigma_1^2 = \sigma_2^2$. From the sample data we work out s_1^2 and s_2^2 under:

Table 10.12

Sample 1			Sample 2		
X_{1i}	$(X_{1i} - \bar{X}_1)$	$(X_{1i} - \bar{X}_1)^2$	X_{2i}	$(X_{2i} - \bar{X}_2)$	$(X_{2i} - \bar{X}_2)^2$
20	−2	4	27	−8	64
16	−6	36	33	−2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	−3	9
22	0	0	34	−1	1
18	−4	16	38	3	9
24	2	4	28	−7	49
25	3	9	41	6	36
19	−3	9	43	8	64
			30	−5	25
			37	2	4
$\Sigma X_{1i} = 220$		$\Sigma(X_{1i} - \bar{X}_1)^2 = 120$	$\Sigma X_{2i} = 420$		$\Sigma(X_{2i} - \bar{X}_2)^2 = 314$
$n_1 = 10$			$n_2 = 12$		

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{220}{10} = 22; \qquad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{420}{12} = 35$$

$\therefore$
$$s_1^2 = \frac{\Sigma (X_{1i} - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10-1} = 13.33$$

nd
$$s_2^2 = \frac{\Sigma (X_{2i} - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12-1} = 28.55$$

Hence,
$$F = \frac{s_2^2}{s_1^2} \qquad (\because s_2^2 > s_1^2)$$

$$= \frac{28.55}{13.33} = 2.14$$

Degrees of freedom in sample 1 = $(n_1 - 1) = 10 - 1 = 9$

Degrees of freedom in sample 2 = $(n_2 - 1) = 12 - 1 = 11$

As the variance of sample 2 is greater variance, hence

$$v_1 = 11;\ v_2 = 9$$

The table value of F at 5 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 3.11 and the tab value of F at 1 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 5.20.

Since the calculated value of $F = 2.14$ which is less than 3.11 and also less than 5.20, the F rat is insignificant at 5 per cent as well as at 1 per cent level of significance and as such we accept t null hypothesis and conclude that samples have been drawn from two populations having the sam variances.

Example 10.30: Given $n_1 = 9;\ n_2 = 8$

$$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 = 184$$

$$\Sigma\left(X_{2i} - \bar{X}_2\right)^2 = 38$$

Apply F-test to judge whether this difference is significant at 5 per cent level.

Solution: We start with the hypothesis that the difference is not significant and hence, $H_0 : \sigma^2 = \sigma$

To test this, we work out the F-ratio as under:

$$F = \frac{\Sigma\,(X_{1i} - \bar{X}_1)^2/(n_1 - 1)}{\Sigma(X_{2i} - \bar{X}_2)^2/(n_2 - 1)}$$

$$= \frac{184/8}{38/7} = \frac{23}{5.43} = 4.25$$

$v_1 = 8$ being the number of d.f. for greater variance.

$v_2 = 7$ being the number of d.f. for smaller variance.

The table value of F at 5 per cent level for $v_1 = 8$ and $v_2 = 7$ is 3.73. Since the calculated value o F is greater than the table value, the F ratio is significant at 5 per cent level. Accordingly we reject H and conclude that the difference is significant.

Example 10.31: Weight of 10 students is as follows:

S. No.	1	2	3	4	5	6	7	8	9	10
Weight (kgs)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the abov sample of 10 students was drawn is equal to 20 kgs? Test this at 5 per cent and 1 per cent level o significance.

lution: First of all we should work out the variance of the sample data or s^2 and the same has been orked out as under:

Table 10.13

S. No.	X (Weight in kgs.)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	−9	81
2	40	−7	49
3	45	−2	04
4	53	+6	36
5	47	+0	00
6	43	−4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
$n = 10$	$\Sigma X_i = 470$		$\Sigma(X_i - \bar{X})^2 = 280$

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{470}{10} = 47 \text{ kgs.}$$

$$s_1 = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}} = \sqrt{31.11}$$

r

$$s_1^2 = 31.11.$$

Let the null hypothesis be $H_0 : \sigma^2 = 31.11$. In order to test this hypothesis we work out the χ^2 alue as under:

$$\chi^2 = \frac{(n-1)s_1^2}{\sigma^2}$$

$$= \frac{31.11}{20}(10-1) = 13.999$$

Degrees of freedom in the given case is $(n-1) = (10-1) = 9$. At 5 per cent level of significance he table value of $\chi^2 = 16.92$ and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both hese values are greater than the calculated value of χ^2 which is 13.999. Hence we accept the null ypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 per ent as also at 1 per cent level of significance. In other words, the sample can be said to have been aken from a population with variance 20 kgs.

Example 10.32: A sample of 10 is drawn randomly from a certain population. The sum of t squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance the population is 5 at 5 per cent level of significance.

Solution: Given information is

$$n = 10$$

$$\Sigma\left(X_i - \overline{X}\right)^2 = 50$$

$$\therefore \qquad s_1^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1} = \frac{50}{9}$$

Take the null hypothesis as $H_0 : \sigma^2 = s_1^2$. In order to test this hypothesis, we work out the χ value as under:

$$\chi^2 = \frac{s_1^2}{\sigma^2}(n-1) = \frac{\frac{50}{9}}{5}(10-1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = (10 – 1) = 9.

The table value of χ^2 at 5 per cent level for 9 d.f. is 16.92. The calculated value of χ^2 is les than this table value, so we accept the null hypothesis and conclude that the variance of the populatio is 5 as given in the question.

FORMULAE SUMMARY

(i) For single mean, single proportion, and single variance:

Testing for single mean:

$$Z_c = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad T_c = \frac{\overline{x} - \mu}{s_1/\sqrt{n}} \sim t_{(n-1)}$$

Testing for single proportion:

$$Z_c = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} \sim N(0, 1)$$

Testing for single variance:

$$\chi_c^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma^2} = \frac{(n-1)s_1^2}{\sigma^2} = \frac{ns^2}{\sigma^2} \sim \chi_{(n-1)}^2$$

(ii) For two means

1. *Independent Samples :* Known σ_1^2, σ_2^2

$$Z_c = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

2. *Independent Samples:* Unknown σ_1^2, σ_2^2

$$T_c = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{(m+n-2)}$$

3. *Related Samples:* Known σ_D^2

$$Z_c = \frac{\bar{d}}{\sigma_D / \sqrt{n}} \sim N(0, 1)$$

4. *Related Samples:* Unknown σ_D^2

$$T_c = \frac{\bar{d}}{s_D / \sqrt{n}} \sim t_{(n-1)}$$

(iii) For two proportions

$$Z_c = \frac{p_1 - p_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \hat{\pi} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

(iv) For two variances

$$F_c = \frac{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{s_1^2(x)}{s_1^2(y)} \sim F(m-1, n-1)$$

PROBLEMS

1. Distinguish between the following:
   (i) Simple hypothesis and composite hypothesis;
   (ii) Null hypothesis and alternative hypothesis;
   (iii) One-tailed test and two-tailed test;
   (iv) Type I error and Type II error;
   (v) Acceptance region and rejection region;
   (vi) Power function and operating characteristic function.
2. What is a hypothesis? What characteristics it must possess in order to be a good research hypothesis A manufacturer considers his production process to be working properly if the mean length of the rod the manufactures is 8.5″. The standard deviation of the rods always runs about 0.26″. Suppose a sample of 64 rods is taken and this gives a mean length of rods equal to 8.6″. What are the null and alternative hypotheses for this problem? Can you infer at 5% level of significance that the process is working properly?
3. The procedure of testing hypothesis requires a researcher to adopt several steps. Describe in brief all such steps.
4. What do you mean by the power of a hypothesis test? How can it be measured? Describe and illustrate by an example.
5. Briefly describe the important parametric tests used in context of testing hypotheses. How such tests differ from non-parametric tests? Explain.
6. Clearly explain how will you test the equality of variances of two normal populations.
7. (a) What is a *t*-test? When it is used and for what purpose(s)? Explain by means of examples.
   (b) Write a brief note on "Sandler's *A*-test" explaining its superiority over *t*-test.
8. Point out the important limitations of tests of hypotheses. What precaution the researcher must take while drawing inferences as per the results of the said tests?
9. A coin is tossed 10,000 times and head turns up 5,195 times. Is the coin unbiased?
10. In some dice throwing experiments, *A* threw dice 41952 times and of these 25145 yielded a 4 or 5 or 6. Is this consistent with the hypothesis that the dice were unbiased?
11. A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out three imperfect articles in a batch of 100. Has the machine improved? Test at 5% level of significance.
12. In two large populations, there are 35% and 30% respectively fair haired people. Is this difference likely to be revealed by simple sample of 1500 and 1000 respectively from the two populations?
13. In a certain association table the following frequencies were obtained:

    $(AB) = 309, (Ab) = 214, (aB) = 132, (ab) = 119.$

    Can the association between *AB* as per the above data can be said to have arisen as a fluctuation of simple sampling?
14. A sample of 900 members is found to have a mean of 3.47 cm. Can it be reasonably regarded as a simple sample from a large population with mean 3.23 cm. and standard deviation 2.31 cm?

5. The means of the two random samples of 1000 and 2000 are 67.5 and 68.0 inches respectively. Can the samples be regarded to have been drawn from the same population of standard deviation 9.5 inches? Test at 5% level of significance.

6. A large corporation uses thousands of light bulbs every year. The brand that has been used in the past has an average life of 1000 hours with a standard deviation of 100 hours. A new brand is offered to the corporation at a price far lower than one they are paying for the old brand. It is decided that they will switch to the new brand unless it is proved with a level of significance of 5% that the new brand has smaller average life than the old brand. A random sample of 100 new brand bulbs is tested yielding an observed sample mean of 985 hours. Assuming that the standard deviation of the new brand is the same as that of the old brand,

   (a) What conclusion should be drawn and what decision should be made?

   (b) What is the probability of accepting the new brand if it has the mean life of 950 hours?

7. Ten students are selected at random from a school and their heights are found to be, in inches, 50, 52, 52, 53, 55, 56, 57, 58, 58 and 59. In the light of these data, discuss the suggestion that the mean height of the students of the school is 54 inches. You may use 5% level of significance (Apply *t*-test as well as *A*-test).

18. In a test given to two groups of students, the marks obtained were as follows:

First Group	18	20	36	50	49	36	34	49	41
Second Group	29	28	26	35	30	44	46		

Examine the significance of difference between mean marks obtained by students of the above two groups. Test at five per cent level of significance.

19. The heights of six randomly chosen sailors are, in inches, 63, 65, 58, 69, 71 and 72. The heights of 10 randomly chosen soldiers are, in inches, 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Do these figures indicate that soldiers are on an average shorter than sailors? Test at 5% level of significance.

20. Ten young recruits were put through a strenuous physical training programme by the army. Their weights (in kg) were recorded before and after with the following results:

Recruit	1	2	3	4	5	6	7	8	9	10
Weight before	127	195	162	170	143	205	168	175	197	136
Weight after	135	200	160	182	147	200	172	186	194	141

Using 5% level of significance, should we conclude that the programme affects the average weight of young recruits (Answer using *t*-test as well as *A*-test)?

21. Suppose a test on the hypotheses $H_0 : \mu = 200$ against $H_1 : \mu > 200$ is done with 1% level of significance, $\sigma = 40$ and $n = 16$.

   (a) What is the probability that the null hypothesis might be accepted when the true mean is really 210? What is the power of the test for $\mu = 210$? How these values of β and $1 - \beta$ change if the test had used 5% level of significance?

   (b) Which is more serious, a Type I and Type II error?

22. The following nine observations were drawn from a normal population:

27 19 20 24 23 29 21 17 27

   (a) Test the null hypothesis $H_0 : \mu = 26$ against the alternative hypothesis $H_1 : \mu \neq 26$. At what level of significance can H_0 be rejected?

   (b) At what level of significance can $H_0 : \mu = 26$ be rejected when tested against $H_1 : \mu < 26$?

23. Suppose that a public corporation has agreed to advertise through a local newspaper if it can b established that the newspaper circulation reaches more than 60% of the corporation's customers What H_0 and H_a should be established for this problem while deciding on the basis of a sample o customers whether or not the corporation should advertise in the local newspaper? If a sample of siz 100 is collected and 1% level of significance is taken, what is the critical value for making a decision whether or not to advertise? Would it make any difference if we take a sample of 25 in place of 100 for ou purpose? If so, explain.

24. Answer using *F*-test whether the following two samples have come from the same population:

Sample 1	17 27 18 25 27 29 27 23 17
Sample 2	16 16 20 16 20 17 15 21

Use 5% level of significance.

25. The following table gives the number of units produced per day by two workers *A* and *B* for a number of days:

A	40 30 38 41 38 35
B	39 38 41 33 32 49 49 34

Should these results be accepted as evidence that *B* is the more stable worker? Use *F*-test at 5% level.

26. A sample of 600 persons selected at random from a large city gives the result that males are 53%. Is there reason to doubt the hypothesis that males and females are in equal numbers in the city? Use 1% level of significance.

27. 12 students were given intensive coaching and 5 tests were conducted in a month. The scores of tests 1 and 5 are given below. Does the score from Test 1 to Test 5 show an improvement? Use 5% level of significance.

No. of students	1	2	3	4	5	6	7	8	9	10	11	12
Marks in 1st test	50	42	51	26	35	42	60	41	70	55	62	38
Marks in 5th test	62	40	61	35	30	52	68	51	84	63	72	50

28. (i) A random sample from 200 villages was taken from Kanpur district and the average population per village was found to be 420 with a standard deviation of 50. Another random sample of 200 villages from the same district gave an average population of 480 per village with a standard deviation of 60. Is the difference between the averages of the two samples statistically significant? Take 1% level of significance.

(ii) The means of the random samples of sizes 9 and 7 are 196.42 and 198.42 respectively. The sums of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the samples be constituted to have been drawn from the same normal population? Use 5% level of significance.

29. A farmer grows crops on two fields *A* and *B*. On *A* he puts Rs. 10 worth of manure per acre and on *B* ₹ 20 worth. The net returns per acre exclusive of the cost of manure on the two fields in the five years are:

Year	1	2	3	4	5
Field A, Rs per acre	34	28	42	37	44
Field B, Rs per acre	36	33	48	38	50

Other things being equal, discuss the question whether it is likely to pay the farmer to continue the more expensive dressing. Test at 5% level of significance.

0. *ABC* Company is considering a site for locating their another plant. The company insists that any location they choose must have an average auto traffic of more than 2000 trucks per day passing the site. They take a traffic sample of 20 days and find an average volume per day of 2140 with standard deviation equal to 100 trucks.

Answer the following:

(i) If $\alpha = 0.05$, should they purchase the site?

(ii) If we assume the population mean to be 2140, what is the β error?

1. An insurance agent claims that the average age of policy holders who have insured through him is less than the average for all agents, which is 30.5 years. A random sample of 100 policy holders who had insured through him gave the following age distribution:

Age	*Number of people*
16–20	12
21–25	22
26–30	20
31–35	30
36–40	16

Test his claim at 5% level of significance.

32. The mean tariff of a hotel room in Las Vegas is said to be \$168 per night. A random sample of 25 hotels resulted in mean tariff \$172.50 with standard deviation \$15.40. Test at the $\alpha = 0.05$ level, if the mean tariff of a hotel room in Las Vegas is more than \$168 per night. A stem-and-leaf display and a normal probability plot indicate the data are approximately normally distributed.

33. An engineer has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for an average of 300 minutes on a single gallon of regular gasoline. Suppose a simple random sample of 50 engines is tested. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance.

34. A political group ABC Party claims that the proportion of voters who support it is more than 0.4. A news channel conducts a survey to test this claim. A sample of 200 people is selected. Out of them only 65 are the supporters of ABC Party. Can you help the news channel in the analysis?

35. A manufacturer believes exactly 8% of its products contain at least one minor flaw. The company wants to test this belief. A sample of 200 products resulted in 33 items have at least one minor flaw. Use a probability of a Type I error of 0.10. Help the company officials.

36. Below are given the gain in weights (in lbs) of pigs fed on two diets A and B.

A:	25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25
B:	44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test if the two diets differ significantly as regards their effect on increase in weight.

37. Samples of two types of electric bulbs were tested for length of life and following data were obtain

	Type A	*Type B*
Sample size	8	7
Sample mean	1234 Hrs	1036 Hrs
Sample s.d.	36 Hrs	40 Hrs

Is the difference in the means sufficient to warrant that type A is superior to type B regarding length life?

38. A certain stimulus administered to each of the 12 patients resulted in the following increase of blo pressure: 5, 2, 8, –1, 3, 0, –2, 1, 5, 0, 4, and 6. Can it be concluded that the stimulus will, in general, accompanied by an increase in blood pressure?

39. An IQ test was given to 5 persons before and after they were trained.

Candidates	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any improvement in IQ after the training program (test at 1% level of significanc

40. Suppose a drug company develops a new drug, designed to prevent colds. The company states th the drug is equally effective for men and women. A sample of 100 women and another sample of 200 m from a population of 100,000 are given the drug. In the samples, 38% of the women caught the cold; an 51% of the men caught the cold. Based on these findings, can we reject the company's claim that t drug is equally effective for men and women?

Based on these findings, can we conclude that the drug is more effective for women than for men? U a 0.05 level of significance.

41. An investment manager manages a portfolio of 250 common stocks. The manager is searching for 'low risk' issue to add to the portfolio, i.e., one with a price variance less than that of the S&P 500 inde Her staff reported that during the last nine quarters the price variance for the S&P 500 index (populatio 1) was 25, and for the last seven quarters the price variance for XYC common (population 2) was Assume that stock prices are normally distributed in the population. Can it be concluded that pric variance for XYC common is mostly smaller?

42. ABC Company purchases steel rods for its projects. Based on previous tests. The quality assuranc manager has recommended purchasing rods from DEF Rods, Inc. (population 1), rather than GHI Ste (population 2), since DEF's rods had less variability in length. Recently, GHI revised its rod shearin operation, and the manager has sampled the outputs from DEF's and GHI's rod manufacturing processe The results for DEF were: sample variance = 0.10, sample size = 8, and, for GHI, the results were samp variance = 0.05, sample size = 10. Do the two variances differ significantly? Assume that rod lengths ar normally distributed in the population.

11

Chapter

hi-Square Tests

the tests of significance discussed in Chapter 10 were based on certain parameters and their mation. These tests were applied under certain assumptions about the population, like normal ribution. These tests can be categorized as parametric tests.

Chi-square distribution is used to obtain confidence interval estimate of unknown population iance. We have already discussed chi-square distribution and interval estimation in Chapter 9. We e also discussed a test of significance for variance using chi-square distribution in the previous pter. In this chapter we use chi-square distribution in some other tests which are not based on any ameter like mean, variance or proportion. Here, the test statistics do not utilize any estimate like nple mean, sample variance or sample proportion. Therefore, the tests to be discussed in this pter can be categorized as non-parametric tests. These tests are (i) test of difference of more n two proportions, (ii) test of independent of attributes, and (iii) test of goodness of fit.

.1 TEST OF DIFFERENCE OF MORE THAN TWO PROPORTIONS

Section 10.11, we discussed the test of difference of two proportions. When we wish to test the erence of more than two proportions, the previous method cannot be used. We below present an mple to understand the idea.

ample 11.1: Suppose assembly elections are announced in three major states. A major political ty wishes to test if the proportions of its supporters in the three states are the same or not. The ty conducts a sample survey of 1000 people in each state and finds that there are 300, 350, and supporters in the sample in the surveyed states. Using these data test if the three proportions are same or not.

Here we wish to test $H_0: \pi_1 = \pi_2 = \pi_3$ (All the proportions are the same) against H_1: Not all $= \pi_2 = \pi_3$ are equal. Note that, we only test if all the proportions are the same or not. In case we clude that all the proportions are not the same, we cannot find which one is different.

We first make the 2 × 3 (2 rows and 3 columns) contingency table or cross tabulation of the giv data as below:

Table 11.1

Supporter	State 1	State 2	State 3
Yes	300	350	425
No	700	650	575

The contingency table has numbers of individuals or frequencies in its cells. We have thı samples with equal sizes; however equal size is not the restriction. Contingency

The test statistic is

$$\chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where, O_{ij} = Observed frequency of the cell in *i*th row and *j*th column

E_{ij} = Expected frequency of the cell in *i*th row and *j*th column

= [Corresponding row total × Corresponding column total]/ Grand total

For the contingency table having *r* rows and *c* columns, the test statistic χ_c^2 follows chi-squa distribution with $(r-1)(c-1)$ d.f. under two conditions (i) total sample size is large (more than 5 and (ii) each cell in the contingency table has expected frequency of at least five. Also rememb that the expected frequencies are grouped in a way so that $\sum O_{ij} = \sum E_{ij}$.

Note that when all the population proportions are the same we expect the O_{ij} and E_{ij} to close which means under null hypothesis the value of computed test statistic χ_c^2 is very small. Thu when the value of computed test statistic χ_c^2 is large, we are more likely to reject the null hypothes That is why the test is considered to be the right tailed test. We reject H_0 in the favor of H_1 $\alpha \times 100\%$ level, if

$$\chi_c^2 > \chi_{(\alpha)}^2$$

Here $\chi_{(\alpha)}^2$ is the critical value obtained using chi square distribution (Table 3 in Appendix) $(r-1)(c-1)$ d.f.

Solution of Example 11.1: We wish to test

$H_0 : \pi_1 = \pi_2 = \pi_3$ (All the proportions are the same) against H_1: Not all π_1, π_2, π_3 are equal.

Calculation of expected frequencies:

Table 11.2

Supporter	State 1	State 2	State 3	
Yes	$O_{11} = 300$ $E_{11} = 358.33$	$O_{12} = 350$ $E_{12} = 358.33$	$O_{13} = 425$ $E_{13} = 358.33$	1075
No	$O_{21} = 700$ $E_{21} = 641.67$	$O_{22} = 650$ $E_{22} = 641.67$	$O_{23} = 575$ $E_{23} = 641.67$	1925
Total	1000	1000	1000	3000

Test Statistic: $$\chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= 9.495127 + 0.193645 + 12.40446 + 5.302397 + 0.108138 + 6.927064$$

$$= 34.43082717$$

Test statistic follows chi square distribution with (2 – 1)(3 – 1) = 2 d.f. Therefore critical value at level of significance is 5.991. Since computed value of test statistic is larger than the critical ue, we reject the null hypothesis at 5% level of significance and conclude that all proportions of pporters of the political party in three states are not the same.

1.2 TEST OF INDEPENDENCE OF ATTRIBUTES

i-square test discussed in previous section can be generalized to test the independence of two ributes of a population. For instance, we may be interested in knowing whether a new medicine is ective in controlling fever or not, χ^2 test will helps us in deciding this issue. In such a situation, we ceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are lependent which means that new medicine is not effective in controlling fever. On this basis we st calculate the expected frequencies and then work out the value of χ^2. If the calculated value of is less than the table value at a certain level of significance for given degrees of freedom, we nclude that null hypothesis stands which means that the two attributes are independent or not sociated (i.e., the new medicine is not effective in controlling the fever). We test if the two attributes nd B of a given population are independent or not. Attribute A is divided into r classes $A_1, A_2, \ldots, A_r$. tribute B is divided into c classes $B_1, B_2, \ldots, B_c$. We are given the data in a $r \times c$ contingency table. re we test:

- H_0: Attributes A and B are independent.
- H_1: Attributes A and B are not independent.

Remember that this test does not measure the degree of relationship or the form of relationship tween two attributes. It is a only a technique of judging the significance of such association or ationship between two attributes.

The test statistic is the same as earlier:

$$\chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where, O_{ij} = Observed frequency of the cell in ith row and jth column

E_{ij} = Expected frequency of the cell in ith row and jth column

= [Corresponding row total × Corresponding column total]/ Grand total

For the contingency table having ***r*** rows and ***c*** columns, the test statistic χ_c^2 follows chi-squ distribution with ***(r−1)(c−1)*** d.f. provided.

(i) total sample size is large (more than 50),

(ii) each cell in the contingency table has expected frequency of at least five,

(iii) $\sum O_{ij} = \sum E_{ij}$.

As earlier, this is also a right tailed test. We reject H_0 in the favor of H_1 at $\alpha \times 100\%$ level, if $\chi_c^2 > \chi$

Example 11.2: The meal plan selected by 200 students is shown below:

Table 11.3

Class Standing	*Number of meals per week*			*Total*
	20/week	*10/week*	*none*	
Fresh	24	32	14	**70**
Sophomore	22	26	12	**60**
Junior	10	14	6	**30**
Senior	14	16	10	**40**
Total	**70**	**88**	**42**	**200**

We wish to test if "Class Standing" is independent of "Number of Meals per Week".

Solution: We wish to test

H_0: "Class Standing" and "Number of Meals per Week" are independent

Against

H_1: "Class Standing" and "Number of Meals per Week" are not independent

Formula of test statistic is $\chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

alculation of expected frequencies
Corresponding row total × Corresponding column total]/ Grand total):

Table 11.4

Class Standing	*Number of meals per week*			*Total*
	20/week	*10/week*	*none*	
Fresh	24.5	30.8	14.7	**70**
Sophomore	21	26.4	12.6	**60**
Junior	10.5	13.2	6.3	**30**
Senior	14	17.6	8.4	**40**
Total	**70**	**88**	**42**	**200**

$$\text{Test Statistic} = \chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 0.709$$

We can see that

(i) Total sample size is 200 (large enough),

(ii) Each expected frequency is higher than 5,

(iii) Sum of observed frequencies = Sum of expected frequencies.

Therefore the distribution of test statistic is chi square distribution with (4 – 1)(3 – 1) = 6 d.f. At % level of significance, for right tailed test, the critical value is 12.592 which is much higher than the omputed value. Therefore, we cannot reject the null hypothesis at 5% level of significance. Thus ve conclude that "Class Standing" and "Number of Meals per Week" may be independent.

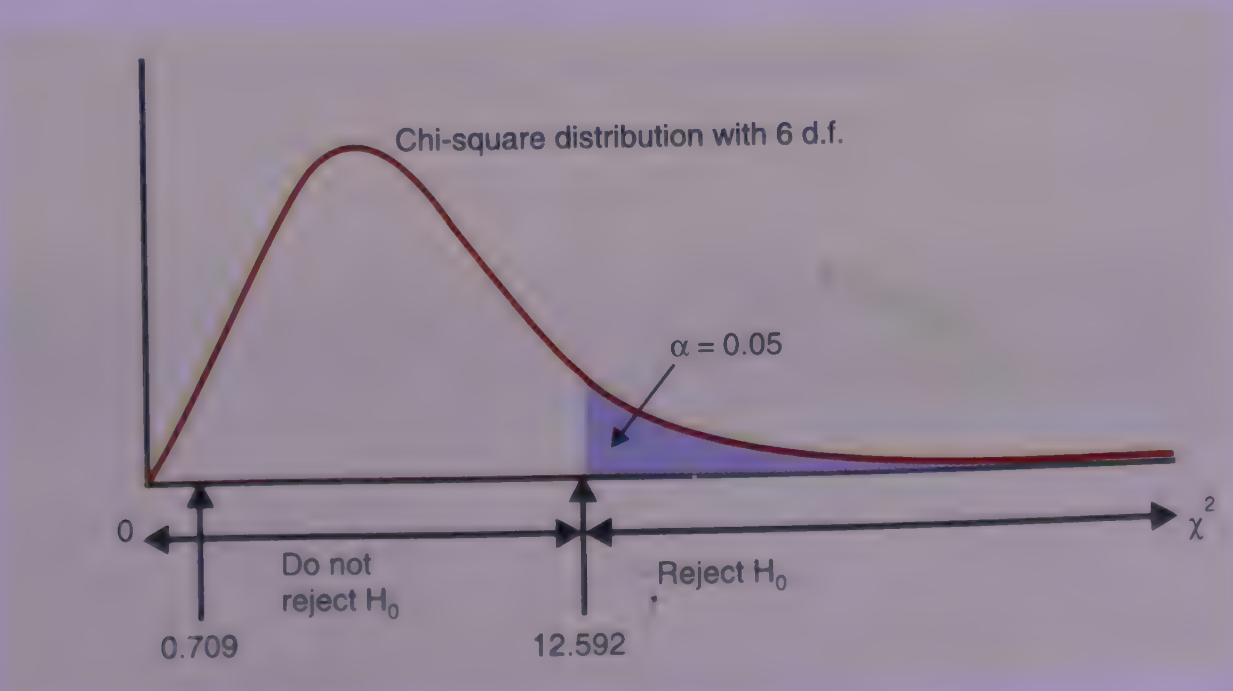


Fig. 11.1

Chi-square distribution is a continuous distribution. It loses its validity if any expected frequen is less than five. In such case, the expected frequency is pooled with the preceding or succeedi frequency. D.f. is reduced by one for one such pooling. This idea is illustrated below.

Example 11.3: Two researchers adopted different sampling techniques while investigating the sa group of customers to find the number of customers falling in different buying-intelligence levels

Table 11.5

Researcher	*No. of customers in each level*				*Total*
	Below average	*Average*	*Above average*	*Genius*	
1	86	60	44	10	200
2	40	33	25	2	100
Total	126	93	69	12	300

Are the two sampling techniques and buying intelligence significantly independent? (Use $\alpha = 0.05$).

Solution: We wish to test

H_0: two sampling techniques and buying intelligence are independent

Against

H_1: two sampling techniques and buying intelligence are not independent

Formula of test statistic is $\chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Calculation of expected frequencies

([Corresponding row total × Corresponding column total]/ Grand total):

Table 11.6

Researcher	*No. of Customers in Each level*				*Total*
	Below Average	*Average*	*Above Average*	*Genius*	
1	84	62	46	8	200
2	42	31	23	4	100
Total	**126**	**93**	**69**	**12**	**300**

Using observed and expected frequencies, we can make the following table:

Table 11.7

O_{ij}	E_{ij}	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
86	84	0.048
60	62	0.064
44	46	0.087
10	8	0.5
40	42	0.095
33	31	0.129
25 2 =27	23 4 =27	0
300	300	0.923

We notice that one expected frequency is 4 which is less than 5, so we add this expected equency with the preceding expected frequency. Corresponding observed frequencies are also lded.

$$\text{Test Statistic} = \chi_c^2 = \sum_{\text{all cells}} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 0.923$$

We can see that

(i) Total sample size is 200 (large enough).

(ii) Each expected frequency is higher than 5, (one expected frequency which was below 5, has been pooled with the preceding frequency).

(iii) Sum of observed frequencies = Sum of expected frequencies.

The distribution of test statistic is chi square distribution with $(2-1)(4-1)-1=2$ d.f. At 5% level f significance, for right tailed test, the critical value is 5.991which is much higher than the computed alue. Therefore, we cannot reject the null hypothesis at 5% level of significance. Thus we conclude at two attributes may be independent.

Note: Table 3 has the critical values for chi-square distribution with degree of freedom 1 to 30. For larger samples N(0, 1) distribution can be used. After obtaining the value of test statistic χ_c^2, apply following formula:

$$Z_c = \sqrt{2\chi_c^2} - \sqrt{2n-1}$$

Distribution of this Z_c is N(0, 1). The computed value of Z_c can be compared with the critical values obtained using N(0, 1) distribution.

11.2.1 Alternative Formula

There is an alternative method of calculating the value of χ^2 in the case of a (2 × 2) table. If write the cell frequencies and marginal totals in case of a (2 × 2) table thus,

a	b	$(a+b)$
c	d	$(c+d)$
$(a+c)$	$(b+d)$	N

then the formula for calculating the value of χ^2 will be stated as follows:

$$\chi^2 = \frac{(ad - bc)^2 \cdot N}{(a+c)(b+d)(a+b)(c+d)}$$

where N means the total frequency, ad means the larger cross product, bc means the smaller cro product and $(a + c)$, $(b + d)$, $(a + b)$, and $(c + d)$ are the marginal totals. The alternative formula rarely used in finding out the value of chi-square as it is not applicable uniformly in all cases but ca be used only in a (2 × 2) contingency table.

11.2.2 Yates' Correction

In a 2 × 2 contingency table, the degree of freedom is (2–1)(2–1)=1. If any expected frequency less than 5, then the use of pooling method will result in chi square with 0 d.f., which is meaningles In this case we apply a correction known as Yates' correction. This consists in adding 0.5 to th observed frequency wish is small and subtracting 0.5 from the other observed frequencies so tha row total and column total are not disturbed. The final formula after adjustment is:

$$\chi_c^2\,(\text{corrected}) = \frac{N(|ad - bc| - 0.5N)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This correction is valid for 2 × 2 contingency table only. The distribution of test statistic is chi square with d.f. = 1.

Example 11.4: The following information is obtained concerning an investigation of 50 ordinar shops of small size:

Table 11.8

	Shops		Total
	In towns	In villages	
Run by Men	12	16	28
Run by Women	3	9	12
Total	15	25	40

Can it be inferred that shops run by women are relatively more in villages than in town? Use 5% vel of significance.

olution: We wish to test

: both attributes are independent (shops run by women are not relatively more in villages than in wn)

gainst

: both attributes are not independent

When we obtain expected frequency, one expected frequency comes to be less than 5. We annot use pooling method as it will result in 0 degree of freedom which is meaningless. So, we use ates' correction. According to which:

$$\chi_c^2\,(\text{corrected}) = \frac{N(|ad-bc|-0.5N)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{40\times(|12\times9-16\times3|-20)^2}{28\times12\times15\times25}$$

At 1 d.f. and 5% level of significance, the critical value is 3.841 which is much higher than the alculated value. Therefore, we cannot reject the null hypothesis at 5% level of significance. Thus, ve conclude that Gender of shopkeepers is independent of shop being in town or in village. We annot say that the shops run by women are relatively more in villages than in town.

1.2.3 Magnitude of Association

ince χ^2 does not by itself provide an estimate of the magnitude of association between two attributes, ny obtained χ^2 value may be converted into Phi coefficient (symbolized as ϕ) for the purpose. In ther words, chi-square tells us about the significance of a relation between variables; it provides no nswer regarding the magnitude of the relation. This can be achieved by computing the Phi coefficient, hich is a non-parametric measure of coefficient of correlation, as under:

$$\varphi = \sqrt{\frac{\chi_c^2}{N}}$$

11.2.4 Coefficient of Contingency

A better measure of relationship or dependency (or contingency) between two attributes is given by coefficient of contingency which is given as

$$C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + N}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

The maximum value of C depends on the order of contingency table. Therefore, we cannot compare the values of C obtained from different types of contingency tables.

11.3 TEST OF GOODNESS OF FIT

Test of goodness of fit is a powerful test given by Karl Pearson. It enables us to find if the deviati of the experiment from theory is just by chance or due to inadequacy of the theory to fit the observ data. The test statistic is exactly the same as discusses earlier, the only difference is in calculating t expected frequencies. To understand the test we illustrate the following example:

Example 11.5: Company *A* has recently conducted aggressive advertising campaigns to maint and possibly increase its share of market for fabric softener. Their main competitor, company ***B*** h 40% of the market. A number of other competitors account for the remaining 15%. To determine the market shares changed after the marketing campaigns, the marketing manager of company solicited the preferences of a random sample of 200 customers. 102 indicated preference for ***A***, for ***B*** and remaining 16 for the others. Can the analyst infer that customer preferences are change (Use a = 0.05).

This test enables us to see how well does the assumed theoretical distribution (such as Binomi distribution, Poisson distribution or Normal distribution) fit to the observed data. When some theoretic distribution is fitted to the given data, we are always interested in knowing as to how well th distribution fits with the observed data. The whole population is divided into a finite number mutually exclusive and exhaustive categories or classes and the data is in terms of number of item or individuals falling in those categories or classes.

The test hypothesis is

H_0: Fit is good or the observed data is drawn from the population having the specified distributio

Against

H_1: Fit is not good the observed data is not drawn from the population having the specifie distribution

The test statistic is:

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where, O_i = Observed frequency of the *i*th class

E_i = Expected frequency of the *i*th class

= Total frequency × Probability of that class

Probability of each class is obtained using the specified theory or distribution. For *k* classes, tes statistic χ_c^2 follows chi-square distribution with $k-1$ d.f. provided:

(i) Total sample size is large (more than 50),

(ii) Each cell in the contingency table has expected frequency of at least five,

(iii) $\sum O_i = \sum E_i$

We reject H_0 in the favor of H_1 at $\alpha \times 100\%$ level, if $\chi_c^2 > \chi_{(\alpha)}^2$.

lution of Example 11.5: We wish to test

: Fit is good. (Data is perfectly fitted to the theory)

: Fit is not good.

Table 11.9

Class	O_i	Market Share (%)	Probability	E_i	$\frac{(O_{ij}-E_i)^2}{E_i}$
Company *A*	102	45	0.45	200 × 0.45 = 90	1.6
Company *B*	82	40	0.40	200 × 0.40 = 80	0.05
Other Companies	16	15	0.15	200 × 0.15 = 30	6.53
	N = 200	**100**	**1**	**200**	**8.18**

The test statistic follows chi square distribution with $3-1=2$ d.f. At 5% level of significance, the ritical value is 5.991 which is smaller than the computed value 8.18. Therefore, at 5% level of gnificance we reject the null hypothesis and conclude that the fit is not good, i.e., the distribution of arket share is changed.

1.3.1 Goodness of Fit Test for Normal Distribution

n many statistical analyses, we assume that the data has normal distribution. We can use test of oodness of fit to examine the normality of a given data. We illustrate following example to understand he method:

xample 11.6: We wish to examine if following data has normal distribution:

5.65	5.44	5.42	5.40	5.53	5.34	5.54	5.45	5.52	5.41
5.57	5.40	5.53	5.54	5.55	5.62	5.56	5.46	5.44	5.51
5.47	5.40	5.47	5.61	5.53	5.32	5.67	5.29	5.49	5.55
5.77	5.57	5.42	5.58	5.58	5.50	5.32	5.50	5.53	5.58
5.61	5.45	5.44	5.25	5.56	5.63	5.5	5.57	5.67	5.36

We wish to examine if the given sample has $N(\mu,\sigma)$ distribution. The parameters μ and σ are ot known. Normal distribution depends on these two parameters. Therefore, before testing the ormality of a given data we need to know these two parameters. In case these parameters are not nown, we can estimate them using the sample observations. Sample mean $\bar{x}$ is used to estimate μ nd sample standard deviation $s_1 = \sqrt{\frac{1}{n-1}(x_i-\bar{x})^2}$ provides a good estimate of σ. Now, we arrange he given data in class intervals. The class frequencies are considered as observed frequencies. In rder to obtain expected frequencies we require the probability of each class. The probability of a class (L, U) is given by P(L < X < U), where X has $N(\mu,\sigma)$ distribution. These probabilities can obtained by first transforming $N(\mu,\sigma)$ to $N(0, 1)$ and then using Table 1 of Appendix. In case μ ar σ are not known, we use their estimates. After obtaining the observed and expected frequencies, use the chi square test statistic. The distribution of test statistic is chi square with d.f. = No. classes – 1 – No. of estimates used.

Solution: We wish to test:

H_0: Given data follows normal distribution.

H_1: Given data does not follow normal distribution.

Since, μ and σ are not known, we estimate them as $\overline{x}=5.5014$ and $s_1=0.10583$, respectively. Calculation of observed and expected frequencies.

Table 11.10

Class Interval	O_i	*Probability*	E_i
Below 5.25	0	P(X < 5.25) = 0.0088	0.438127
5.25 – 5.35	5	P(5.25 < X < 5.35) = 0.0675	3.375546
5.35 – 5.45	10	P(5.35 < X < 5.45) = 0.2373	11.86612
5.45 – 5.55	17	P(5.45 < X < 5.55) = 0.3634	18.16841
5.55 – 5.65	14	P(5.55 < X < 5.65) = 0.2429	12.14483
5.65 – 5.75	3	P(5.65 < X < 5.75) = 0.0707	3.536422
5.75 – 5.85	1	P(5.75 < X < 5.85) = 0.0089	0.445844
Above 5.85	0	P(X > 5.85) = 0.0005	0.024697
	50		50

We can notice that some expected frequencies are below 5. We add all those expected frequencies and corresponding observed frequencies.

Table 11.11

O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
9	7.82064	0.17785
10	11.8661	0.29348
14	12.1448	0.28338
17	18.1684	0.07514
		$\chi_c^2 = 0.82985$

The test statistic $\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 0.82985$

The test statistic follows chi square distribution with d.f. = Final number of classes – 1 – No. of imates = 4 – 1 – 2 =1.

At 5% level of significance, the critical value is 3.841 which is larger than the computed value 2985. Therefore, at 5% level of significance we cannot reject the null hypothesis. Thus there is no tistical evidence against normality of the given data.

Note: We can take different class intervals. But the minimum number of classes (after ooling) should be such that the d.f. is a positive integer. In case one parameter is known out of and σ, we need to estimate only the unknown one. In such a case d.f. = Final number of lasses – 1 – No. of estimates = $k - 1 - 1$, for k classes.

Note: Like normal distribution, we can also test the goodness of fit for other distributions also.

1.4 CAUTION IN USING CHI SQUARE TESTS

e chi-square test is no doubt a most frequently used test, but its correct application is equally an hill task. It should be borne in mind that the test is to be applied only when the individual observations sample are independent which means that the occurrence of one individual observation (event) s no effect upon the occurrence of any other observation (event) in the sample under consideration. mall theoretical frequencies, if these occur in certain groups, should be dealt with under special re. The other possible reasons concerning the improper application or misuse of this test can be) neglect of frequencies of non-occurrence; (ii) failure to equalise the sum of observed and the sum the expected frequencies; (iii) wrong determination of the degrees of freedom; (iv) wrong omputations, and the like. The researcher while applying this test must remain careful about all ese things and must thoroughly understand the rationale of this important test before using it and rawing inferences in respect of his hypothesis.

lore Solved Examples

xample: 11.7: A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

olution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining ny one of the six numbers is 1/6 and as such the expected frequency of any one number coming pward is $132 \times 1/6 = 22$. Now we can write the observed frequencies along with expected frequencies nd work out the value of χ^2 as follows:

Table 11.12

No. turned up	Observed frequency O_i	Expected frequency E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
1	16	22	–6	36	36/22
2	20	22	–2	4	4/22
3	25	22	3	9	9/22
4	14	22	–8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\therefore \qquad \Sigma\,[(O_i - E_i)^2/E_i] = 9.$$

Hence, the calculated value of $\chi^2 = 9$.

$\because$ Degrees of freedom in the given problem is

$(n - 1) = (6 - 1) = 5.$

The table value* of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.07 Comparing calculated and table values of χ^2, we find that calculated value is less than the tab value and as such could have arisen due to fluctuations of sampling. The result, thus, supports th hypothesis and it can be concluded that the die is unbiased.

Example 11.8: Find the value of χ^2 for the following information:

Class	*A*	*B*	*C*	*D*	*E*
Observed frequency	8	29	44	15	4
Theoretical (or expected) frequency	7	24	38	24	7

Solution: Since some of the frequencies less than 10, we shall first re-group the given data a follows and then will work out the value of χ^2:

Table 11.13

Class	Observed frequency O_i	Expected frequency E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
A and *B*	(8 + 29) = 37	(7 + 24) = 31	6	36/31
C	44	38	6	36/38
D and *E*	(15 + 4) = 19	(24 + 7) = 31	–12	144/31

$$\therefore \qquad \chi^2 = \Sigma\frac{(O_i - E_i)^2}{E_i} = 6.76 \text{ app.}$$

---

* Table No. 3 showing some critical values of χ^2 for specified degrees of freedom has been given in Appendix at the end of the book.

ample: 11.9: Genetic theory states that children having one parent of blood type A and the other blood type B will always be of one of three types, A, AB, B and that the proportion of three types ll on an average be as 1 : 2 : 1. A report states that out of 300 children having one A parent and B rent, 30 per cent were found to be types A, 45 per cent per cent type AB and remainder type B.

st the hypothesis by χ^2 test.

lution: The observed frequencies of type A, AB and B is given in the question are 90, 135 and 75 spectively.

The expected frequencies of type A, AB and B (as per the genetic theory) should have been 75, 0 and 75 respectively.

We now calculate the value of χ^2 as follows:

Table 11.14

Type	*Observed frequency* O_i	*Expected frequency* E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
A	90	75	15	225	225/75 = 3
AB	135	150	–15	225	225/150 = 1.5
B	75	75	0	0	0/75 = 0

$$\therefore \quad \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 3 + 1.5 + 0 = 4.5$$

$$\because \quad \text{d.f.} = (n - 1) = (3 - 1) = 2.$$

Table value of χ^2 for 2 d.f. at 5 per cent level of significance is 5.991.

The calculated value of χ^2 is 4.5 which is less than the table value and hence can be ascribed o have taken place because of chance. This supports the theoretical hypothesis of the genetic theory hat on an average type A, AB and B stand in the proportion of 1 : 2 : 1.

xample: 11.10: The table given below shows the data obtained during outbreak of smallpox:

	Attacked	*Not attacked*	*Total*
Vaccinated	31	469	500
Not vaccinated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result vith the help of χ^2 at 5 per cent level of significance.

Solution: Let us take the hypothesis that vaccination is not effective in preventing the attack from smallpox i.e., vaccination and attack are independent. On the basis of this hypothesis, the expected frequency corresponding to the number of persons vaccinated and attacked would be:

$$\text{Expectation of } (AB) = \frac{(A) \times (B)}{N}$$

when A represents vaccination and B represents attack.

$$\therefore \quad (A) = 500$$

$$(B) = 216$$

$$N = 2000$$

$$\text{Expectation of } (AB) = \frac{500 \times 216}{2000} = 54$$

Now using the expectation of (AB), we can write the table of expected values as follows:

Table 11.15

	Attacked: B	Not attacked: b	Total
Vaccinated: A	$(AB) = 54$	$(Ab) = 446$	500
Not vaccinated: a	$(aB) = 162$	$(ab) = 1338$	1500
Total	216	1784	2000

Table 11.16: Calculation of Chi-Square

Group	Observed frequency O_{ij}	Expected frequency E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$(O_{ij} - E_{ij})^2/E_{ij}$
AB	31	54	–23	529	529/54 = 9.796
Ab	469	446	+23	529	529/44 = 1.186
aB	158	162	+23	529	529/162 = 3.265
ab	1315	1338	–23	529	529/1338 = 0.395

$$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = 14.642$$

$\because$ Degrees of freedom in this case = $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.

The table value of χ^2 for 1 degree of freedom at 5 per cent level of significance is 3.841. Th calculated value of χ^2 is much higher than this table value and hence the result of the experimen does not support the hypothesis. We can, thus, conclude that vaccination is effective in preventin the attack from smallpox.

ample: 11.11: Two research workers classified some people in income groups on the basis of mpling studies. Their results are as follows:

Investigators	*Income groups*			*Total*
	Poor	*Middle*	*Rich*	
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Show that the sampling technique of at least one research worker is defective.

olution: Let us take the hypothesis that the sampling techniques adopted by research workers are milar (i.e., there is no difference between the techniques adopted by research workers). This being , the expectation of *A* investigator classifying the people in

(i) Poor income group $= \dfrac{200 \times 300}{500} = 120$

(ii) Middle income group $= \dfrac{200 \times 150}{500} = 60$

(iii) Rich income group $= \dfrac{200 \times 50}{500} = 20$

Similarly the expectation of *B* investigator classifying the people in

(i) Poor income group $= \dfrac{300 \times 300}{500} = 180$

(ii) Middle income group $= \dfrac{300 \times 150}{500} = 90$

(iii) Rich income group $= \dfrac{300 \times 50}{500} = 30$

We can now calculate value of χ^2 as follows:

Table 11.17

Groups	*Observed frequency* O_{ij}	*Expected frequency* E_{ij}	$O_{ij} - E_{ij}$	$(O_{ij} - E_{ij})^2/E_{ij}$
Investigator A				
classifies people as poor	160	120	40	1600/120 = 13.33
classifies people as middle class people	30	60	–30	900/60 = 15.00
classifies people as rich	10	20	–10	100/20 = 5.00
Investigator B				
classifies people as poor	140	180	–40	1600/180 = 8.88
classifies people as middle class people	120	90	30	900/90 = 10.00
classifies people as rich	40	30	10	100/30 = 3.33

Hence,l $$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = 55.54$$

$$\because \text{ Degrees of freedom } = (c - 1)(r - 1)$$
$$= (3 - 1)(2 - 1) = 2.$$

The table value of χ^2 for two degrees of freedom at 5 per cent level of significance is 5.9 The calculated value of χ^2 is much higher than this table value which means that the calcula value cannot be said to have arisen just because of chance. It is significant. Hence, the hypothe does not hold good. This means that the sampling techniques adopted by two investigators differ a are not similar. Naturally, then the technique of one must be superior than that of the other.

Example: 11.12: Eight coins were tossed 256 times and the following results were obtained:

Numbers of heads	0	1	2	3	4	5	6	7	8
Frequency	2	6	30	52	67	56	32	10	1

Are the coins biased? Use χ^2 test.

Solution: Let us take the hypothesis that the coins are not biased. If that is so, the probability of a one coin falling with head upward is 1/2 and with tail upward is 1/2 and it remains the same whatev be the number of throws. In such a case the expected values of getting 0, 1, 2, … heads in a sing throw in 256 throws of eight coins will be worked out as follows*.

Table 11.18

Events or No. of heads	*Expected frequencies*
0	$^8C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^8 \times 256 = 1$
1	$^8C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^7 \times 256 = 8$
2	$^8C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^6 \times 256 = 28$
3	$^8C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^5 \times 256 = 56$
4	$^8C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^4 \times 256 = 70$

---

*The probabilities of random variable i.e., various possible events have been worked out on the binomial principle viz through the expansion of $(p + q)^n$ where $p = 1/2$ and $q = 1/2$ and $n = 8$ in the given case. The expansion of the ter "$^nC_r\, p^r q^{n-r}$" has given the required probabilities which have been multiplied by 256 to obtain the expected frequencies.

Events or No. of heads	Expected frequencies
5	${}^8C_5\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^3 \times 256 = 56$
6	${}^8C_6\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 \times 256 = 28$
7	${}^8C_7\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^1 \times 256 = 8$
8	${}^8C_8\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^0 \times 256 = 1$

The value of χ^2 can be worked out as follows:

Table 11.19

No. of heads	Observed frequency O_i	Expected frequency E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	2	1	1	1/1 = 1.00
1	6	8	–2	4/8 = 0.50
2	30	28	2	4/28 = 0.14
3	52	56	–4	16/56 = 0.29
4	67	70	–3	9/70 = 0.13
5	56	56	0	0/56 = 0.00
6	32	28	4	16/28 = 0.57
7	10	8	2	4/8 = 0.50
8	1	1	0	0/1 = 0.00

$$\therefore \quad \chi^2 = \Sigma\frac{(O_i - E_i)^2}{E_i} = 3.13$$

$\therefore$ Degrees of freedom = $(n - 1) = (9 - 1) = 8$

The table value of χ^2 for eight degrees of freedom at 5 per cent level of significance is 15.507.

he calculated value of χ^2 is much less than this table and hence it is insignificant and can be scribed due to fluctuations of sampling. The result, thus, supports the hypothesis and we may say at the coins are not biased.

Example 11.13: Following are the number of aircraft accidents during the various days of t week. Test at % level of significance if the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thu	Fri	Sat	Sun
No. of Acc.	14	16	8	12	11	9	14

Solution: We wish to test:

H_0: Fit is good. (Data follows discrete uniform distribution)

H_1: Fit is not good.

In case of (discrete) uniform distribution, the probability of each class is the same which is 1 here.

The test statistic is:

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Therefore,

Table 11.20

								Total
O_i	14	16	8	12	11	9	14	84
E_i	12	12	12	12	12	12	12	84
$\frac{(O_i - E_i)^2}{E_i}$	0.3333	1.3333	1.3333	0.0000	0.0833	0.7500	0.3333	4.1667

The test statistic follows chi square distribution with $7-1=6$ d.f. At 5% level of significance. th critical value is 12.592 which is larger than the computed value 4.1667. Therefore, at 5% level o significance we cannot reject the null hypothesis. Thus there is no statistical evidence that the accident are not uniformly distributed over the week.

Example 11.14: A survey of 320 families with 5 children each revealed the following distribution:

No. of boys	5	4	3	2	1	0
No. of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution: We wish to test

H_0: Male and female births are equally probable

H_1: Male and female births are not equally probable

Success: A randomly chosen child is a girl

X = No. of girl children in a family with 5 children.

Under H_0: $X \sim B(5, 1/2)$

$P(X = 0) = {}^5C_0(1/2)^0\,(1 - 1/2)^{5-0} = {}^5C_0(1/2)^5 = 0.03125$

$P(X = 1) = {}^5C_1(1/2)^5 = 0.15625$

$P(X = 2) = {}^5C_2(1/2)^5 = 0.31250$

$P(X = 3) = {}^5C_3(1/2)^5 = 0.31250$

$P(X = 4) = {}^5C_4(1/2)^5 = 0.15625$

$P(X = 5) = {}^5C_5(1/2)^5 = 0.03125$

The test statistic is:

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Therefore,

Table 11.21

O_i	E_i	$\frac{(O_i-E_i)^2}{E_i}$
14	$320 \times P(X=0) = 10$	1.6
56	$320 \times P(X=1) = 50$	0.72
110	$320 \times P(X=2)\ 100$	1
88	$320 \times P(X=3) = 100$	1.44
40	$320 \times P(X=4) = 50$	2
12	$320 \times P(X=5) = 10$	0.4
N = 320	**320**	$\chi_c^2 = 7.16$

The test statistic follows chi square distribution with $6 - 1 = 5$ d.f. At 5% level of significance, the tical value is 11.070 which is larger than the computed value 7.16. Therefore, at 5% level of gnificance we cannot reject the null hypothesis. Thus there is no statistical evidence that male and nale births are not equally probable.

PROBLEMS

1. What is Chi-square text? Explain its significance in statistical analysis.
2. Write short notes on the following:
    (i) Additive property of Chi-square.
    (ii) Chi-square as a test of 'goodness of fit'.
    (iii) Precautions in applying Chi-square test.
    (iv) Conditions for applying Chi-square test.
3. An experiment was conducted to test the efficacy of chloromycetin in checking typhoid. In a certai hospital chloromycetin was given to 285 out of the 392 patients suffering from typhoid. The number typhoid cases were as follows:

	Typhoid	*No typhoid*	*Total*
Chloromycetin	35	250	285
No chloromycetin	50	57	107
Total	85	307	392

With the help of χ^2, test the effectiveness of chloromycetin in checking typhoid.

(The χ^2 value at 5 per cent level of significance for one degree of freedom is 3.841).

4. On the basis of information given below about the treatment of 200 patients suffering from a disease state whether the new treatment is comparatively superior to the conventional treatment.

Treatment	*No. of patients*	
	Favourable response	*No. response*
New	60	20
conventional	70	50

For drawing your inference, use the value of χ^2 for one degree of freedom at the 5 per cent level o significance, viz., 3.84.

5. 200 digits were chosen at random from a set of tables. The frequencies of the digits were:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Calculate χ^2.

6. Five dice were thrown 96 times and the number of times 4, 5, or 6 was thrown were

Number of dice throwing

4, 5 or 6	5	4	3	2	1	0
Frequency	8	18	35	24	10	1

Find the value of Chi-square.

. Find Chi-square from the following information:

Condition of child	*Condition of home*		*Total*
	Clean	*Dirty*	
Clean	70	50	120
Fairly clean	80	20	100
Dirty	35	45	80
Total	185	115	300

State whether the two attributes viz., condition of home and condition of child are independent (Use Chi-square test for the purpose).

8. In a certain cross the types represented by XY, Xy, xY and xy are expected to occur in a 9 : 5 : 4 : 2 ratio. The actual frequencies were:

XY	Xy	xY	xy
180	110	60	50

Test the goodness of fit of observation to theory.

9. The normal rate of infection for a certain disease in cattle is known to be 50 per cent. In an experiment with seven animals injected with a new vaccine it was found that none of the animals caught infection. Can the evidence be regarded as conclusive (at 1 per cent level of significance) to prove the value of the new vaccine?

0. Result of throwing die were recorded as follows:

Number falling upwards	1	2	3	4	5	6
Frequency	27	33	31	29	30	24

Is the die unbiased? Answer on the basis of Chi-square test.

11. The Theory predicts the proportion of beans, in the four groups A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the number in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory? Apply χ^2 test.

2. You are given a sample of 150 observations classified by two attributes A and B as follows:

	A_1	A_2	A_3	Total
B_1	40	25	15	80
B_2	11	26	8	45
B_3	9	9	7	25
Total	60	60	30	150

Use the χ^2 test to examine whether A and B are associated.

13. A survey of 320 families with five children each revealed the following distribution:

No. of boys	5	4	3	2	1	0
No. of girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

Is this distribution consistent with the hypothesis that male and female births are equally probabl
Apply Chi-square test.

14. What is Yates' correction? Find the value of Chi-square applying Yates' correction to the following da

	Passed	*Failed*	*Total*
Day classes	10	20	30
Evening classes	4	66	70
Total	14	86	100

Also state whether the association, if any, between passing in the examination and studying in d classes is significant using Chi-square test.

15. (a) 1000 babies were born during a certain week in a city of which 600 were boys and 400 girls. U χ^2 test to examine the correctness of the hypothesis that the sex-ratio is 1 : 1 in newly born babie

(b) The percentage of smokers in a certain city was 90. A random sample of 100 persons was selecte in which 85 persons were found to be smokers. Is the sample proportion significantly differe from the proportion of smokers in the city? Answer on the basis of Chi-square test.

16. A college is running post-graduate classes in five subjects with equal number of students. The tot number of absentees in these five classes is 75. Test the hypothesis that these classes are alike i absenteeism if the actual absentees in each are as follows:

History = 19
Philosophy = 18
Economics = 15
Commerce = 12
Chemistry = 11

17. The number of automobile accidents per week in a certain community were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were the same during th 10 week period under consideration?

18. A certain chemical plant processes sea water to collect sodium chloride and magnesium. From scientifi analysis, sea water is known to contain sodium chloride, magnesium and other elements in the ratio o 62 : 4 : 34. A sample of 200 tons of sea water has resulted in 130 tons of sodium chloride and 6 tons o magnesium. Are these data consistent with the scientific model at 5 per cent level of significance?

An oil company has explored three different areas for possible oil reserves. The results of the test were as given below:

	Area			*Total*
	A	*B*	*C*	
Strikes Dry holes	7 10	10 18	8 9	25 37
Total number of test wells	17	28	17	62

Do the three areas have the same potential, at the 10 per cent level of significance?

. While conducting an air traffic study, a record was made of the number of aircraft arrivals, at a certain airport, during 250 half hour time intervals. The following tables gives the observed number of periods in which there were 0, 1, 2, 3, 4, or more arrivals as well as the expected number of such periods if arrivals per half hour have a Poisson distribution $\lambda = 2$. Does this Poisson distribution describe the observed arrivals at 5 per cent level of significance.

Number of observed arrivals (per half hour)	*Number of periods observed*	*Number of periods expected (Poisson, $\lambda = 2$)*
0	47	34
1	56	68
2	71	68
3	44	45
4 or more	32	35

1. A marketing researcher interested in the business publication reading habits of purchasing agents has assembled the following data:

Business Publication Preferences (First Choice Mentions)

Business publication	*Frequency of first choice*
A	35
B	30
C	45
D	55

(i) Test the null hypothesis ($\alpha = 0.05$) that there are no differences among frequencies of first choice of tested publications.

(ii) If the choice of *A* and *C* and that of *B* and *D* are aggregated, test the null hypothesis at $\alpha = 0.05$ that there are no differences.

22. A group of 150 College students were asked to indicate their most liked film star from among six diffe well known film actors viz., *A*, *B*, *C*, *D*, *E* and *F* in order to ascertain their relative popularity. observed frequency data were as follows:

Actors	*A*	*B*	*C*	*D*	*E*	*F*	Total
Frequencies	24	20	32	25	28	21	150

Test at 5 per cent whether all actors are equally popular.

23. For the data in question 12, find the coefficient of contingency to measure the magnitude of relations between *A* and *B*.

24. (a) What purpose is served by calculating the Phi coefficient (ϕ)? Explain.

(b) If $\chi^2 = 16$ and $N = 4$, find the value of Phi coefficient.

25. Suppose a die is tossed 60 times. It is observed that – 1 comes 8 times, 2 comes 6 times, 3 comes 12 tim 4 comes 8 times, 5 comes 12 times, and 6 comes 14 times. Test at 5% level of significance if the die is f

26. Weekly demand of a product is assumed to be normally distributed. Use a goodness of fit test and t following data to test this assumption. Use $\alpha = 0.10$, sample mean = 24.5, sample standard deviation =

18	20	22	27	22	26	25	25	27	25
25	22	27	25	24	25	28	24	25	26
26	23	20	24	26	31	29	28	19	21

nalysis of Variance

12

Chapter

lysis of variance (abbreviated as ANOVA) is an extremely useful technique concerning researches e fields of economics, biology, education, psychology, sociology, business/industry and in researches everal other disciplines. This technique is used when multiple sample cases are involved. As ed earlier, the significance of the difference between the means of two samples can be judged ough either z-test or the t-test, but the difficulty arises when we happen to examine the significance he difference amongst more than two sample means at the same time. The ANOVA technique bles us to perform this simultaneous test and as such is considered to be an important tool of lysis in the hands of a researcher. Using this technique, one can draw inferences about whether samples have been drawn from populations having the same mean.

2.1 THE ANOVA TECHNIQUE

e ANOVA technique is important in the context of all those situations where we want to compare re than two populations such as in comparing the yield of crop from several varieties of seeds, the oline mileage of four automobiles, the smoking habits of five groups of university students and so In such circumstances one generally does not want to consider all possible combinations of two pulations at a time for that would require a great number of tests before we would be able to arrive decision. This would also consume lot of time and money, and even then certain relationships y be left unidentified (particularly the interaction effects). Therefore, one quite often utilizes the OVA technique and through it investigates the differences among the means of all the populations ultaneously.

Professor R.A. Fisher was the first man to use the term ‘Variance’ and, in fact, it was he who eloped a very elaborate theory concerning ANOVA, explaining its usefulness in practical field. er on Professor Snedecor and many others contributed to the development of this technique. OVA is essentially a procedure for testing the difference among different groups of data for mogeneity. “The essence of ANOVA is that the total amount of variation in a set of data is broken wn into two types, that amount which can be attributed to chance and that amount which can be ributed to specified causes.” There may be variation between samples and also within sample ms. ANOVA consists in splitting the variance for analytical purposes. Hence, it is a method of analysing the variance to which a response is subject into its various components correspondi various sources of variation. Through this technique one can explain whether various varieti seeds or fertilizers or soils differ significantly so that a policy decision could be taken accordi concerning a particular variety in the context of agriculture researches. Similarly, the differenc various types of feed prepared for a particular class of animal or various types of drugs manufact for curing a specific disease may be studied and judged to be significant or not through the applica of ANOVA technique. Likewise, a manager of a big concern can analyse the performance of var salesmen of his concern in order to know whether their performances differ significantly.

Thus, in general, through ANOVA technique one can, investigate any number of factors w are hypothesized or said to influence the dependent variable. One may as well investigate differences amongst various categories within each of these factors which may have a large num of possible values. If we take only one factor and investigate the differences amongst its var categories having numerous possible values, we are said to use one-way ANOVA and in case investigate two factors at the same time, then we use two-way ANOVA. In two way ANOVA interaction or inter-relationship of two factors affecting the values of a variable can as wel studied for better decisions.

12.2 THE BASIC PRINCIPLE OF ANOVA

The basic principle of ANOVA is to test for differences among the means of the population examining the amount of variation within each of these samples, relative to the amount of varia between the samples. In terms of variation within the given population, it is assumed that the observa (x_{ij}) differ from the mean of this population only because of random effects i.e., there are influen on (x_{ij}) which are unexplainable, whereas in examining differences between populations we assu that the difference between the mean of the *j*th population and the grand mean is attributable to w is called a 'specific factor' or what is technically described as treatment effect. Thus while us ANOVA, we assume that each of the samples is drawn from a normal population and that each these populations has the same variance. We also assume that all factors other than the one or m being tested are effectively controlled. This, in other words, means that we assume the absence many factors that might affect our conclusions concerning the factor(s) to be studied.

The total variance in the joint sample is partitioned into two parts (i) between samples varian and (ii) within samples variance. Between samples variance is due to different treatments, wh within samples variance is due to the random unexplained disturbance. Using these two varianc we define the test statistic as:

$$F_C = \frac{\text{between samples variance}}{\text{within samples variance}}.$$

Using this method we wish to test

H_0 : all population means are the same (or effects of all treatments are the same)

Against

H_1 : all population means are not the same (or effects of all treatments are not the sam

Between sample variance is large when the effects of all the treatments are different. In such a e the computed F_c is large and we are likely to reject the null hypothesis. Therefore, this test is a ht tailed test.

2.3 ONE WAY ANOVA

der the one-way ANOVA, we consider only one factor and then observe that the reason for said ctor to be important is that several possible types of samples can occur within that factor. We then termine if there are differences within that factor. The technique involves the following steps:

(i) Obtain the mean of each sample i.e., obtain

$$\bar{x}_1,\ \bar{x}_2,\ \bar{x}_3,\ \ldots,\ \bar{x}_k$$

when there are k samples.

(ii) Work out the mean of the sample means as follows:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \ldots + \bar{x}_k}{\text{No. of samples } (k)}$$

(iii) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the number of items in the corresponding sample, and then obtain their total. This is known as the sum of squares for variance between the samples (or *SS* between). Symbolically, this can be written as:

$$SS \text{ between} = n_1 (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 (\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k (\bar{x}_k - \bar{\bar{x}})^2$$

(iv) Divide the result of the (iii) by the degrees of freedom between the samples to obtain variance or mean square (*MS*) between samples. Symbolically, this can be written as:

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

where $(k - 1)$ represents degrees of freedom (d.f.) between samples.

(v) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain their total. This total is known as the sum of squares for variance within samples (or *SS* within). Symbolically this can be written as:

$$SS \text{ within} = \sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2i} - \bar{x}_2)^2 + \ldots + \sum (x_{ki} - \bar{x}_k)^2$$

$$i = 1, 2, 3, \ldots$$

(vi) Divide the result of (v) by the degrees of freedom within samples to obtain the variance or mean square (*MS*) within samples. Symbolically, this can be written as:

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

where $(n - k)$ represents degrees of freedom within samples,

n = Total number of items in all the samples i.e., $n_1 + n_2 + \ldots + n_k$

k = Number of samples.

(vii) For a check, the sum of squares of deviations for total variance can also be worked out adding the squares of deviations when the deviations for the individual items in all t samples have been taken from the mean of the sample means. Symbolically, this can written as:

$$SS \text{ for total variance} = \sum (x_{ij} - \bar{\bar{x}})^2 \qquad i = 1, 2, 3, \ldots \text{ and } j = 1, 2, 3, \ldots$$

This total should be equal to the total of the result of the (iii) and (v) explained above i.e

$$SS \text{ for total variance} = SS \text{ between} + SS \text{ within.}$$

The degrees of freedom for total variance will be equal to the number of items in a samples minus one i.e., $(n - 1)$. The degrees of freedom for between and within must ac up to the degrees of freedom for total variance i.e.,

$$(n - 1) = (k - 1) + (n - k)$$

(viii) Finally, F-ratio may be worked out as under:

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

This F-ratio works as the test statistic and follows snedico's F-distribution with $(k - 1$ $(n-k)$ degree of freedom. A discussed the test is a right tailed test. Therefore we reject th null hypothesis that all the population means (or the effects of all the treatments) are th same at given level of significance when the computed value of F-ratio is greater than th critical value which can be obtained from Tables 4(a) and 4(b).

12.3.1 Analysis of Variance Table

For the sake of convenience the information obtained through various steps stated above can be pu as under:

Table 12.1: Analysis of Variance Table for One-way ANOVA
(There are *k* samples having in all *n* items)

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)	Mean square (MS) (This is SS divided by d.f.) and is an estimation of variance to be used in F-ratio	F-ratio
Between samples or categories	SS between	$(k-1)$	MS between $= \frac{SS \text{ between}}{(k-1)}$	$\frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	SS within ...	$(n-k)$	MS within $= \frac{SS \text{ within}}{(n-k)}$	
Total	$\sum (x_{ij} - \bar{\bar{x}})^2$ $i = 1, 2, \dots$ $j = 1, 2, \dots$	$(n-1)$		

2.3.2 Short-cut Method for One-way ANOVA

NOVA can be performed by following the short-cut method which is usually used in practice since e same happens to be a very convenient method, particularly when means of the samples and/or ean of the sample means happen to be non-integer values. The various steps involved in the short- it method are as under:

(i) Take the total of the values of individual items in all the samples i.e., work out $\sum x_{ij}$

$i = 1, 2, 3, \dots$

$j = 1, 2, 3, \dots$

and call it as T.

(ii) Work out the correction factor as under:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

(iii) Find out the square of all the item values one by one and then take its total. Subtract correction factor from this total and the result is the sum of squares for total varia Symbolically, we can write:

$$\text{Total } SS = \sum x_{ij}^2 - \frac{(T)^2}{n} \qquad i = 1, 2, 3, \ldots \text{ and } j = 1, 2, 3, \ldots$$

(iv) Obtain the square of each sample total $(T_j)^2$ and divide such square value of each sam by the number of items in the concerning sample and take the total of the result th obtained. Subtract the correction factor from this total and the result is the sum of squa for variance between the samples. Symbolically, we can write:

$$SS \text{ between} = \Sigma\frac{(T_j)^2}{n_j} - \frac{(T)^2}{n} \qquad j = 1, 2, 3, \ldots$$

where subscript j represents different samples or categories.

(v) The sum of squares within the samples can be found out by subtracting the result of (step from the result of (iii) step stated above and can be written as under:

$$SS \text{ within} = \left\{\sum x_{ij}^2 - \frac{(T)^2}{n}\right\} - \left\{\Sigma\frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}\right\}$$

$$= \sum x_{ij}^2 - \sum\frac{(T_j)^2}{n_j}$$

After doing all this, the table of ANOVA can be set up in the same way as explained earlier.

Note: F-ratio does not change if we add/subtract or multiply/divide all the observations by a constant. This property can be utilized in order to simplify the calculation when observation ar large numbers.

Example 12.1: Set up an analysis of variance table for the following per acre production data f three varieties of wheat, each grown on 4 plots and state if the variety differences are significant

Table 12.2

Per acre production data			
Plot of land	*Variety of wheat*		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

ution: We wish to test

All wheat varieties are the same or effects of different varieties on wheat production are the
e

All varieties are not the same

We can solve the problem by the direct method or by short-cut method, but in each case we shall
the same result. We try below both the methods.

Solution through direct method: First we calculate the mean of each of these samples:

$$\bar{x}_1=\frac{6+7+3+8}{4}=6$$

$$\bar{x}_2=\frac{5+5+3+7}{4}=5$$

$$\bar{x}_3=\frac{5+4+3+4}{4}=4$$

Mean of the sample means or

$$\bar{\bar{x}}=\frac{\bar{x}_1+\bar{x}_2+\bar{x}_3}{k}$$

$$=\frac{6+5+4}{3}=5$$

Now we work out *SS* between and *SS* within samples:

$$\begin{aligned} SS\text{ between} &= n_1(\bar{x}_1-\bar{\bar{x}})^2+n_2(\bar{x}_2-\bar{\bar{x}})^2+n_3(\bar{x}_3-\bar{\bar{x}})^2 \\ &= 4(6-5)^2+4(5-5)^2+4(4-5)^2 \\ &= 4+0+4 \\ &= 8 \end{aligned}$$

$$\begin{aligned} SS\text{ within} &= \sum(x_{1i}-\bar{x}_1)^2+\sum(x_{2i}-\bar{x}_2)^2+\sum(x_{3i}-\bar{x}_3)^2;\quad i=1,2,3,4 \\ &= \{(6-6)^2+(7-6)^2+(3-6)^2+(8-6)^2\} \\ &\quad+\{(5-5)^2+(5-5)^2+(3-5)^2+(7-5)^2\} \\ &\quad+\{(5-4)^2+(4-4)^2+(3-4)^2+(4-4)^2\} \\ &= \{0+1+9+4\}+\{0+0+4+4\}+\{1+0+1+0\} \\ &= 14+8+2 \\ &= 24 \end{aligned}$$

$$\begin{aligned} SS\text{ for total variance} &= \sum(x_{ij}-\bar{\bar{x}})^2 \qquad i=1,2,3\ldots\text{ and } j=1,2,3\ldots \\ &= (6-5)^2+(7-5)^2+(3-5)^2+(8-5)^2 \\ &\quad+(5-5)^2+(5-5)^2+(3-5)^2 \\ &\quad+(7-5)^2+(5-5)^2+(4-5)^2 \\ &\quad+(3-5)^2+(4-5)^2 \end{aligned}$$

$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1$$

$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

$$= 8 + 24 = 32$$

We can now set up the ANOVA table for this problem:

Table 12.3

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*Critical value F(at (from the F-tabl*
Between sample	8	(3 – 1) = 2	8/2 = 4.00	4.00/2.67 = 1.5	*F*(2, 9) = 4.26
Within sample	24	(12 – 3) = 9	24/9 = 2.67		
Total	32	(12 – 1) = 11			

The above table shows that the calculated value of *F* is 1.5 which is less than the table value 4.26 at 5% level with d.f. being $v_1 = 2$ and $v_2 = 9$ and hence could have arisen due to chance. T analysis supports the null-hypothesis of no difference is means. We may, therefore, conclude that difference in wheat output due to varieties is insignificant and is just a matter of chance.

Solution through short-cut method: In this case we first take the total of all the individ values of *n* items and call it as *T*.

T in the given case = 60

and $n = 12$

Hence, the correction factor = $(T)^2/n = 60 \times 60/12 = 300$. Now total *SS*, *SS* between and within can be worked out as under:

$$\text{Total } SS = \sum x_{ij}^2 - \frac{(T)^2}{n}; \quad i = 1, 2, 3, \ldots \text{ and } j = 1, 2, 3, \ldots$$

$$= (6)^2 + (7)^2 + (3)^2 + (8)^2 + (5)^2 + (5)^2 + (3)^2 + (7)^2 + (5)^2$$

$$+ (4)^2 + (3)^2 + (4)^2 - \left(\frac{60 \times 60}{12}\right)$$

$$= 332 - 300 = 32$$

$$SS \text{ between} = \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$$

$$= \left(\frac{24 \times 24}{4}\right) + \left(\frac{20 \times 20}{4}\right) + \left(\frac{16 \times 16}{4}\right) - \left(\frac{60 \times 60}{12}\right)$$

$$= 144 + 100 + 64 - 300$$

$$= 8$$

$$SS \text{ within} = \Sigma\, x_{ij}^2 - \Sigma \frac{(T_j)^2}{n_j}$$

$$= 332 - 308$$

$$= 24$$

It may be noted that we get exactly the same result as we had obtained in the case of direct ethod. From now onwards we can set up ANOVA table and interpret F-ratio in the same manner we have already done under the direct method.

2.3.3 Critical Difference

, on the basis of ANOVA, we reject H_0, i.e., if there is significant difference among various eatment effects, then we would be interested to find out which pair(s) of treatments differ significantly. or this purpose, we use Scheffe's Test. We obtain the critical difference for each pair of treatments sing the formula

Critical Difference between ith and jth treatments

$$= \sqrt{\left[\text{MS within}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \times (k-1) \times F_{k-1,n-k,\alpha}\right]}$$

Here MS within is mean squares within samples, n_i and n_j are corresponding sample sizes, k is umber of treatments, $F_{k-1,n-k,\alpha}$ is the critical value for right tailed test, at $\alpha \times 100\%$ level of ignificance and $(k-1, n-k)$ d.f., obtained using F distribution. We conclude that at $\alpha \times 100\%$ level f significance, ith and jth treatment effects differ significantly if the difference between orresponding sample means is higher than the critical difference.

xample 12.2: You want to see if three different golf clubs yield different distances. You randomly elect five measurements from trials on an automated driving machine for each club.

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235.	197
237	227	206
251	216	204

At the 0.05 significance level, is there a difference in mean distance?

Solution: We wish to test

H_0: No significance difference in the distances by three golf clubs

H_1: The distances by three golf clubs are significantly different

As in the previous example, we make the following ANOVA table:

Table 12.4

Source of variation	Sum of squares	Degree of freedom	Mean sum of squares	Varianc ratio
Treatments (Between Groups)	4716.4	2	2358.2	25.275
Error (Within Groups)	1119.6	12	93.3	
Total	5836	14		

Test Statistic given by variance ratio follows ***F*** distribution with d.f. (2, 12). At 5% level c significance, the critical value is 3.89 which is smaller than the computed value .Therefore, we reje H_0 at 5% level of significance.

After rejecting the null hypothesis we would like to know which pairs are different. To know th we obtain the critical differences for each pair of treatments.

We can see that $n_1 = n_2 = n_3 = 5$, $k = 3$, MS within = 93.3, and F critical value at 5% level is 3.8 Since all sample sizes are the same, therefore all critical differences will be the same, given as

$$\sqrt{\left[\text{MS within}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\times(k-1)\times F_{k-1,n-k,\alpha}\right]}=\sqrt{\left[93.3\left(\frac{1}{5}+\frac{1}{5}\right)\times(3-1)\times 3.89\right]}=17.039$$

Sample means due to Club 1, Club 2, and Club 3 are 249.2, 226, and 205.8 respectively. We mak the following table:

Table 12.5

Pair of treatments	Difference of sample means	Critical difference	Conclusion at 5% level
Club 1 and Club 2	23.2	17.0396	Club 1 and Club 2 differ significantly
Club 2 and Club 3	20.2	17.0396	Club 2 and Club 3 differ significantly
Club 3 and Club 1	43.4	17.0396	Club 3 and Club 1 differ significantly

Thus all clubs have different effect on distance.

2.4 TWO WAY ANOVA

vo way ANOVA technique is used when the data are classified on the basis of two factors. For ample, the agricultural output may be classified on the basis of different varieties of seeds and also the basis of different varieties of fertilizers used. A business firm may have its sales data classified n the basis of different salesmen and also on the basis of sales in different regions. In a factory, the rious units of a product produced during a certain period may be classified on the basis of different rieties of machines used and also on the basis of different grades of labour. Such a two way design ay have repeated measurements of each factor or may not have repeated values. The ANOVA chnique is little different in case of repeated measurements where we also compute the interaction riation. We shall now explain the two way ANOVA technique in the context of both the said esigns with the help of examples.

2.4.1 One Observation per Cell

s we do not have repeated values, we cannot directly compute the sum of squares within samples s we had done in the case of one-way ANOVA. Therefore, we have to calculate this residual or ror variation by subtraction, once we have calculated (just on the same lines as we did in the case f one-way ANOVA) the sum of squares for total variance and for variance between varieties of ne factor as also for variance between varieties of the other factor.

The various steps involved are as follows:

(i) Take the total of observations in all the samples and call it T.

(ii) Work out the correction factor as under:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

(iii) Find out the square of all the item values (or their coded values as the case may be) one by one and then take its total. Subtract the correction factor from this total to obtain the sum of squares of deviations for total variance. Symbolically, we can write it as:

Sum of squares of deviations for total variance or total SS

$$= \sum x_{ij}^2 - \frac{(T)^2}{n}$$

(iv) Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns or (SS between columns).

(v) Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the number of items in the corresponding row and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (or SS between rows).

(vi) Sum of squares of deviations for residual or error variance can be obtained as

SS for residual or error variance = Total *SS* – *SS* between columns – *SS* between ro

(vii) Degrees of freedom (d.f.) can be worked out as under:

d.f. for total variance $= (c \, . \, r - 1)$
d.f. for variance between columns $= (c - 1)$
d.f. for variance between rows $= (r - 1)$
d.f. for residual variance $= (c - 1)\,(r - 1)$

where c = Number of columns
r = Number of rows

(viii) ANOVA table can be set up in the usual fashion as shown below:

Table 12.6: Analysis of Variance Table for Two-way ANOVA

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean square (MS)*	*F-ratio*
Between columns treatment	$\sum\frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c-1)$	$\frac{SS \text{ between columns}}{(c-1)}$	$\frac{MS \text{ between columns}}{MS \text{ residual}}$
Between rows treatment	$\sum\frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r-1)$	$\frac{SS \text{ between rows}}{(r-1)}$	$\frac{MS \text{ between rows}}{MS \text{ residual}}$
Residual or error	Total *SS* – (*SS* between columns + *SS* between rows)	$(c-1)(r-1)$	$\frac{SS \text{ residual}}{(c-1)\,(r-1)}$	
Total	$\sum x_{ij}^2 - \frac{(T)^2}{n}$	$(c.r-1)$		

We here wish to test

(i) H_{OC} : The varieties of first factor have the same effect

Against

H_{ac} : The varieties of the first factor are significantly different

(ii) H_{OR} : The varieties of second factor have the same effect

Against

H_{aR}: The varieties of second factor are significantly different

For hypotheses (i), the test statistic is

$$F_{\text{column}} = \frac{\text{MS between columns}}{\text{MS residuals}}$$

It follows F distribution with $((c-1, (c-1)\,(r-1))$ d.f.

For hypotheses (ii), the test statistic is

$$F_{\text{row}} = \frac{\text{MS between rows}}{\text{MS residuals}}$$

It follows F distribution with $((r-1), (c-1)(r-1))$ d.f.

Thus, *MS* residual or the residual variance provides the basis for the *F*-ratios concerning variation etween columns treatment and between rows treatment. *MS* residual is always due to the fluctuations f sampling, and hence serves as the basis for the significance test. Both the *F*-ratios are compared with their corresponding table values, for given degrees of freedom at a specified level of significance, as usual and if it is found that the calculated *F*-ratio concerning variation between columns is greater han its table value, then the difference among columns means is considered significant. Similarly, the F-ratio concerning variation between rows can be interpreted.

Example 12.3: Set up an analysis of variance table for the following two-way design results:

Per Acre Production Data of Wheat

(*in metric tonnes*)

Varieties of fertilizers	*Varieties of seeds*		
	A	*B*	*C*
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Also state whether variety differences are significant at 5% level.

Solution: As the given problem is a two-way design of experiment without repeated values, we shall adopt all the above stated steps while setting up the ANOVA table as is illustrated on the following page.

Table 12.7: Computations for Two-way ANOVA (in a design without repeated values)

Step (i) $T = 60, n = 12, \therefore$ Correction factor $= \frac{(T)^2}{n} = \frac{60 \times 60}{12} = 300$

Step (ii) Total *SS* $= (36+25+25+49+25+16+9+9+9+64+49+16) - \left(\frac{60 \times 60}{12}\right)$

$= 332 - 300$

$= 32$

Step (iii) *SS* between columns treatment $= \left[\frac{24 \times 24}{4} + \frac{20 \times 20}{4} + \frac{16 \times 16}{4}\right] - \left[\frac{60 \times 60}{12}\right]$

$= 144 + 100 + 64 - 300$

$= 8$

Step (iv) *SS* between rows treatment $= \left[\frac{16 \times 16}{3} + \frac{16 \times 16}{3} + \frac{9 \times 9}{3} + \frac{19 \times 19}{3}\right] - \left[\frac{60 \times 60}{12}\right]$

$= 85.33 + 85.33 + 27.00 + 120.33 - 300$

$= 18$

Step (v) *SS* residual or error = Total *SS* – (*SS* between columns + *SS* between rows)

$= 32 - (8 + 18) = 6$

Table 12.8: The ANOVA Table

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (or the tables values)*
Between columns (i.e., between varieties of seeds)	8	$(3-1)=2$	$8/2=4$	$4/1=4$	$F(2,6)=5.14$
Between rows (i.e., between varieties of fertilizers)	18	$(4-1)=3$	$18/3=6$	$6/1=6$	$F(3,6)=4.76$
Residual or error	6	$(3-1)\times(4-1)=6$	$6/6=1$		
Total	32	$(3\times4)-1=11$			

ANOVA table can be set up for the given problem as shown in Table 12.8.

From the said ANOVA table, we find that differences concerning varieties of seeds are insignificant at 5% level as the calculated *F*-ratio of 4 is less than the table value of 5.14, but the variety differences concerning fertilizers are significant as the calculated *F*-ratio of 6 is more than its table value of 4.76.

12.4.2 More than One Observations Per Cell

In case of a two-way design with repeated measurements for all of the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure we can calculate the sum of squares and degrees of freedom in the same way as we had worked out the sum of squares for variance within samples in the case of one-way ANOVA. Total *SS*, *SS* between columns and *SS* between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as '*interaction variation*' (Interaction is the measure of inter relationship among the two different classifications). After making all these computations, ANOVA table can be set up for drawing inferences. We illustrate the same with an example.

Example 12.4: Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:

Amount of Blood Pressure Reduction in Millimeters of Mercury

Group of people	*Drug*		
	X	*Y*	*Z*
A	14 15	10 9	11 11
B	12 11	7 8	10 11
C	10 11	11 11	8 7

Do the drugs act differently?

Are the different groups of people affected differently?

Is the interaction term significant?

Answer the above questions taking a significant level of 5%.

Solution: We first make all the required computations as shown below:

We can set up ANOVA table shown in Table 12.10.

Table 12.9: Computations for Two-way ANOVA (in design with repeated values)

Step (i) $T = 187$, $n = 18$, thus, the correction factor $= \dfrac{187 \times 187}{18} = 1942.72$

Step (ii)

$$\text{Total } SS = [(14)^2 + (15)^2 + (12)^2 + (11)^2 + (10)^2 + (11)^2 + (10)^2 + (9)^2 + (7)^2 + (8)^2 + (11)^2 + (11)^2 + (11)^2 + (11)^2 + (10)^2 + (11)^2 + (8)^2 + (7)^2] - \left[\frac{(187)^2}{18}\right]$$

$$= (2019 - 1942.72)$$

$$= 76.28$$

Step (iii)

$$SS \text{ between columns (i.e., between drugs)} = \left[\frac{73 \times 73}{6} + \frac{56 \times 56}{6} + \frac{58 \times 58}{6}\right] - \left[\frac{(187)^2}{18}\right]$$

$$= 888.16 + 522.66 + 560.67 - 1942.72$$

$$= 28.77$$

Step (iv)

$$SS \text{ between rows (i.e., between people)} = \left[\frac{70 \times 70}{6} + \frac{59 \times 59}{6} + \frac{58 \times 58}{6}\right] - \left[\frac{(187)^2}{18}\right]$$

$$= 816.67 + 580.16 + 560.67 - 1942.72$$

$$= 14.78$$

Step (v)

$$\begin{aligned} SS \text{ within samples} = {} & (14 - 14.5)^2 + (15 - 14.5)^2 + (10 - 9.5)^2 + (9 - 9.5)^2 + (11 - 11)^2 + (11 - 11)^2 \\ & + (12 - 11.5)^2 + (11 - 11.5)^2 + (7 - 7.5)^2 + (8 - 7.5)^2 \\ & + (10 - 10.5)^2 + (11 - 10.5)^2 + (10 - 10.5)^2 + (11 - 10.5)^2 \\ & + (11 - 11)^2 + (11 - 11)^2 + (8 - 7.5)^2 + (7 - 7.5)^2 \end{aligned}$$

$$= 3.50$$

Step (vi)

$$SS \text{ for interaction variation} = 76.28 - [28.77 + 14.78 + 3.50]$$

$$= 29.23$$

Table 12.10: The ANOVA Table

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*Critical at 5%*
Between columns (i.e., between drugs)	28.77	$(3-1)=2$	$\frac{28.77}{2}$ $=14.385$	$\frac{14.385}{0.389}$ $=36.9$	$F(2, 9)=4.26$
Between rows (i.e., between people)	14.78	$(3-1)=2$	$\frac{14.78}{2}$ $=7.390$	$\frac{7.390}{0.389}$ $=19.0$	$F(2, 9)=4.26$
Interaction	29.23	4	$\frac{29.23}{4}$ $=7.308$	$\frac{7.308}{0.389}$	$F(4, 9)=3.63$
Within samples (Error)	3.50	$17-2-2$ $-4=9$	$\frac{3.50}{9}$ $=0.389$		
Total	76.28	$(18-1)=17$			

Here,

d.f. for interaction = d.f. for colums × d.f. for rows.

d.f. for error = d.f. for total
– d.f. for column
– d.f. for rows
– d.f. for interactions

The above table shows that all the three *F*-ratios are significant at 5% level which means that the drugs act differently, different groups of people are affected differently and the interaction term is significant.

Graphic method of studying interaction in a two-way design: Interaction can be studied in a two-way design with repeated measurements through graphic method also. For such a graph we shall select one of the factors to be used as the *X*-axis. Then we plot the averages for all the samples on the graph and connect the averages for each variety of the other factor by a distinct mark (or a coloured line). If the connecting lines do not cross over each other, then the graph indicates that there is no interaction, but if the lines do cross, they indicate definite interaction or inter-relation between the two factors. Let us draw such a graph for the data of illustration 3 of this chapter to see whether there is any interaction between the two factors viz., the drugs and the groups of people.

Graph of the averages for amount of blood pressure reduction in millimeters of mercury for different drugs and different groups of people.

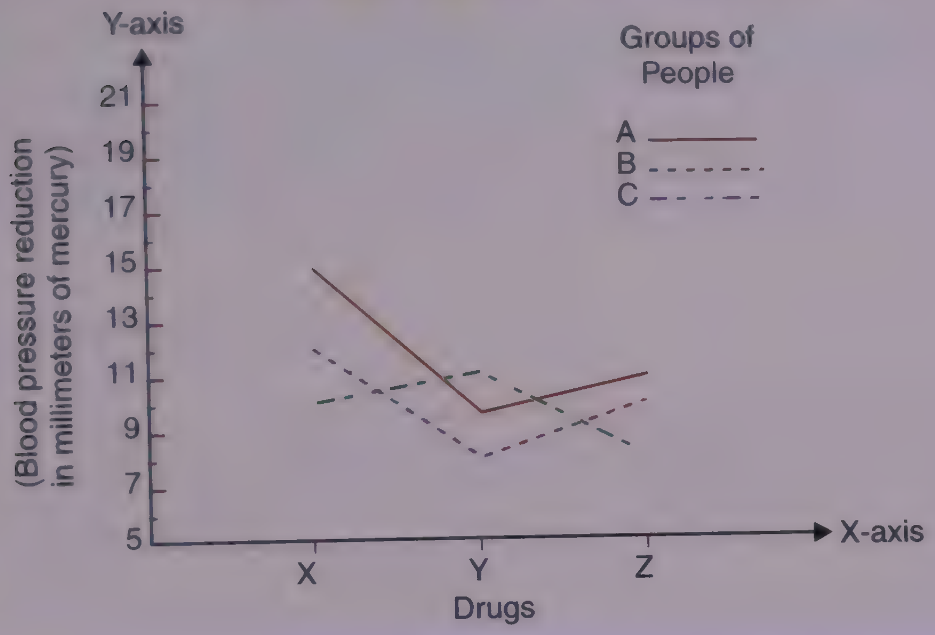


Fig. 12.1

The graph can alternatively be drawn by taking different group of people on X-axis and drawing ines for various drugs through the averages.

The graph indicates that there is a significant interaction because the different connecting lines or groups of people do cross over each other. We find that A and B are affected very similarly, but C is affected differently. The highest reduction in blood pressure in case of C is with drug Y and the owest reduction is with drug Z, whereas the highest reduction in blood pressure in case of A and B is with drug X and the lowest reduction is with drug Y. Thus, there is definite inter-relation between the drugs and the groups of people and one cannot make any strong statements about drugs unless he also qualifies his conclusions by stating which group of people he is dealing with. In such a situation, performing F-tests is meaningless. But if the lines do not cross over each other (and remain more or less identical), then there is no interaction or the interaction is not considered a significantly large value, in which case the researcher should proceed to test the main effects, drugs and people in the given case, as stated earlier.

12.5 LATIN-SQUARE DESIGN

Latin-square design is an experimental design used frequently in agricultural research. In such a design the treatments are so allocated among the plots that no treatment occurs, more than once in any one row or any one column. The ANOVA technique in case of Latin-square design remains more or less the same as we have already stated in case of a two-way design, excepting the fact that the variance is splitted into four parts as under:

(i) variance between columns;
(ii) variance between rows;
(iii) variance between varieties;
(iv) residual variance.

All these above stated variances are worked out as under:

Table 12.11

Variance between columns or *MS* between columns	$= \dfrac{\Sigma \dfrac{(T_j)^2}{n_j} - \dfrac{(T)^2}{n}}{(c-1)} = \dfrac{SS \text{ between columns}}{\text{d.f.}}$
Variance between rows or *MS* between rows	$= \dfrac{\Sigma \dfrac{(T_i)^2}{n_i} - \dfrac{(T)^2}{n}}{(r-1)} = \dfrac{SS \text{ between rows}}{\text{d.f.}}$
Variance between varieties or *MS* between varieties	$= \dfrac{\Sigma \dfrac{(T_v)^2}{n_v} - \dfrac{(T)^2}{n}}{(v-1)} = \dfrac{SS \text{ between varieties}}{\text{d.f.}}$
Residual or error variance or *MS* residual	$= \dfrac{\text{Total } SS - (SS \text{ between columns} + SS \text{ between rows} + SS \text{ between varieties})}{(c-1)(c-2)}$ where total $SS = \Sigma (x_{ij})^2 - \dfrac{(T)^2}{n}$ c = number of columns r = number of rows v = number of varieties, also $c = r = v$

Example 12.5: Analyse and interpret the following statistics concerning output of wheat per field obtained as a result of experiment conducted to test four varieties of wheat viz., *A*, *B*, *C* and *D* under a Latin-square design.

C 25	B 23	A 20	D 20
A 19	D 19	C 21	B 18
B 19	A 14	D 17	C 20
D 17	C 20	B 21	A 15

Solution: To simplify, we subtract 20 from the figures given in each of the small squares and obtain the coded figures as under:

	Columns				Row totals
Rows	*C* 5	*B* 3	*A* 0	*D* 0	8
	A –1	*D* –1	*C* 1	*B* –2	– 2
	B –1	*A* – 6	*D* – 3	*C* 0	–10
	D – 3	*C* 0	*B* 1	*A* – 5	–7
Column totals	0	– 4	–1	–7	T = –12

Fig. 12.2 (a)

Squaring these coded figures in various columns and rows we have:

		Squares of figures				Sum of squares
		Columns				
		1	2	3	4	
	1	C 25	B 9	A 0	D 0	34
Rows	2	A 1	D 1	C 1	B 4	7
	3	B 1	A 36	D 9	C 0	46
	4	D 9	C 0	B 1	A 25	35
Sum of squares		36	46	11	29	T = 122

Fig. 12.2 (b)

$$\text{Correction factor} = \frac{(T)^2}{n} = \frac{(-12)\,(-12)}{16} = 9$$

$$SS \text{ for total variance} = \sum (x_{ij})^2 - \frac{(T)^2}{n} = 122 - 9 = 113$$

$$SS \text{ for variance between columns} = \Sigma \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$$

$$= \left\{ \frac{(0)^2}{4} + \frac{(-4)^2}{4} + \frac{(-1)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{66}{4} - 9 = 7.5$$

SS for variance between rows

$$= \Sigma \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n} \left\{ \frac{(8)^2}{4} + \frac{(-3)^2}{4} + \frac{(-10)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{222}{4} - 9 = 46.5$$

SS for variance between varieties would be worked out as under:

For finding *SS* for variance between varieties, we would first rearrange the data in the followin form:

Table 12.12

Varieties of wheat	*Yield in different parts of field*				*Total (T_v)*
	I	II	III	IV	
A	–1	–6	0	–5	–12
B	–1	3	1	–2	1
C	5	0	1	0	6
D	–3	–1	–3	0	–7

Now we can work out *SS* for variance between varieties as under:

$$SS \text{ for variance between varieties} = \Sigma \frac{(T_v)^2}{n_v} - \frac{(T)^2}{n}$$

$$= \left\{ \frac{(-12)^2}{4} + \frac{(1)^2}{4} + \frac{(6)^2}{4} + \frac{(-7)^2}{4} \right\} - 9$$

$$= \frac{230}{4} - 9 = 48.5$$

∴ Sum of squares for residual variance will work out to

$$113 - (7.5 + 46.5 + 48.5) = 10.50$$

d.f. for variance between columns $= (c-1) = (4-1) = 3$

d.f. for variance between rows $= (r-1) = (4-1) = 3$

d.f. for variance between varieties $= (v-1) = (4-1) = 3$

d.f. for total variance $= (n-1) = (16-1) = 15$

d.f. for residual variance $= (c-1)(c-2) = (4-1)(4-2) = 6$

ANOVA table can now be set up as shown below:

Table 12.13: The ANOVA Table in Latin-square Design

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*Critical at 5%*
Between columns	7.50	3	$\frac{7.50}{3} = 2.50$	$\frac{2.50}{1.75} = 1.43$	$F(3,6) = 4.76$
Between rows	46.50	3	$\frac{46.50}{3} = 15.50$	$\frac{15.50}{1.75} = 8.85$	$F(3,6) = 4.76$
Between varieties	48.50	3	$\frac{48.50}{3} = 16.17$	$\frac{16.17}{1.75} = 9.24$	$F(3,6) = 4.76$
Residual or error	10.50	6	$\frac{10.50}{6} = 1.75$		
Total	113.00	15			

The above table shows that variance between rows and variance between varieties are significant nd not due to chance factor at 5% level of significance as the calculated values of the said two ariances are 8.85 and 9.24 respectively which are greater than the table value of 4.76. But variance etween columns is insignificant and is due to chance because the calculated value of 1.43 is less han the table value of 4.76.

12.6 ANALYSIS OF CO-VARIANCE (ANOCOVA)

he object of experimental design in general happens to be to ensure that the results observed may e attributed to the treatment variable and to no other causal circumstances. For instance, the esearcher studying one independent variable, *X*, may wish to control the influence of some uncontrolled ariable (sometimes called the covariate or the concomitant variables), *Z*, which is known to be orrelated with the dependent variable, *Y*, then he should use the technique of analysis of covariance or a valid evaluation of the outcome of the experiment. "In psychology and education primary nterest in the analysis of covariance rests in its use as a procedure for the statistical control of an ncontrolled variable."

While applying the ANOCOVA technique, the influence of uncontrolled variable is usually remov by simple linear regression* method and the residual sums of squares are used to provide varian estimates which in turn are used to make tests of significance. In other words, covariance analy consists in subtracting from each individual score (Y_i) that portion of it Y_i' that is predictable fr uncontrolled variable (Z_i) and then computing the usual analysis of variance on the resulti ($Y - Y'$)'s, of course making the due adjustment to the degrees of freedom because of the fact th estimation using regression method required loss of degrees of freedom.

12.7 ASSUMPTIONS IN ANOCOVA

The ANOCOVA technique requires one to assume that there is some sort of relationship betwe the dependent variable and the uncontrolled variable. We also assume that this form of relationship the same in the various treatment groups. Other assumptions are:

(i) Various treatment groups are selected at random from the population.
(ii) The groups are homogeneous in variability.
(iii) The relationship is linear and is similar from group to group.

The *short-cut method for ANOCOVA* can be explained by means of an example as show below:

***Example 12.6*:** The following are paired observations for three experimental groups:

Group I		*Group II*		*Group III*	
X	*Y*	*X*	*Y*	*X*	*Y*
7	2	15	8	30	15
6	5	24	12	35	16
9	7	25	15	32	20
15	9	19	18	38	24
12	10	31	19	40	30

Y is the covariate (or concomitant) variable. Calculate the adjusted total, within groups an between groups, sums of squares on *X* and test the significance of differences between the adjuste means on *X* by using the appropriate *F*-ratio. Also calculate the adjusted means on *X*.

Solution: We apply the technique of analysis of covariance and work out the related measures a under:

Table 12.14

Group I		*Group II*		*Group III*	
X	*Y*	*X*	*Y*	*X*	*Y*
7	2	15	8	30	15
6	5	24	12	35	16

	Group I		Group II		Group III	
	X	Y	X	Y	X	Y
	9	7	25	15	32	20
	15	9	19	18	38	24
	12	10	31	19	40	30
Total	49	33	114	72	175	105
Mean	9.80	6.60	22.80	14.40	35.00	21.00

$$\sum X = 49 + 114 + 175 = 338$$

$$\text{Correction factor for } X = \frac{(\sum X)^2}{N} = 7616.27$$

$$\sum Y = 33 + 72 + 105 = 210$$

$$\text{Correction factor for } Y = \frac{(\sum Y)^2}{N} = 2940$$

$$\sum X^2 = 9476 \quad \sum Y^2 = 3734 \quad \sum XY = 5838$$

$$\text{Correction factor for } XY = \frac{\sum X \cdot \sum Y}{N} = 4732$$

Hence, total SS for $X = \sum X^2$ – Correction factor for X

$$= 9476 - 7616.27 = 1859.73$$

$$SS \text{ between for } X = \left\{\frac{(49)^2}{5} + \frac{(114)^2}{5} + \frac{(175)^2}{5}\right\} - \{\text{Correction factor for } X\}$$

$$= (480.2 + 2599.2 + 6125) - (7616.27)$$

$$= 1588.13$$

$$SS \text{ within for } X = (\text{Total } SS \text{ for } X) - (SS \text{ between for } X)$$

$$= (1859.73) - (1588.13) = 271.60$$

Similarly we work out the following values in respect of Y

$$\text{total } SS \text{ for } Y = \sum Y^2 - \text{Correction factor for } Y$$

$$= 3734 - 2940 = 794$$

$$SS \text{ between for } Y = \left\{\frac{(33)^2}{5} + \frac{(72)^2}{5} + \frac{(105)^2}{5}\right\} - \{\text{Correction factor for } Y\}$$

$$= (217.8 + 1036.8 + 2205) - (2940) = 519.6$$

$$SS \text{ within for } Y = (\text{Total } SS \text{ for } Y) - (SS \text{ between for } Y)$$

$$= (794) - (519.6) = 274.4$$

Then, we work out the following values in respect of both X and Y

$$\text{Total sum of product of } XY = \Sigma XY - \text{Correction factor for } XY$$

$$= 5838 - 4732 = 1106$$

$$SS \text{ between for } XY = \left\{\frac{(49)(33)}{5} + \frac{(114)(72)}{5} + \frac{(175)(105)}{5}\right\} - \{\text{Correction factor for } XY\}$$

$$= (323.4 + 1641.6 + 3675) - (4732) = 908$$

$$SS \text{ within for } XY = (\text{Total sum of product}) - (SS \text{ between for } XY)$$

$$= (1106) - (908) = 198$$

ANOVA table for X, Y and XY can now be set up as shown below:

ANOVA Table for X, Y and XY

Source	*d.f.*	*SS for X*	*SS for Y*	*Sum of product XY*
Between groups	2	1588.13	519.60	908
Within groups	12	$E_{XX} = 271.60$	$E_{YY} = 274.40$	$E_{XY} = 198$
Total	14	$T_{XX} = 1859.73$	$T_{YY} = 794.00$	$T_{XY} = 1106$

$$\text{Adjusted total } SS = T_{XX} - \frac{(T_{XY})^2}{T_{YY}}$$

$$= 1859.73 - \frac{(1106)^2}{794}$$

$$= (1859.73) - (1540.60)$$

$$= 319.13$$

$$\text{Adjusted } SS \text{ within group} = E_{XX} - \frac{(E_{XY})^2}{E_{YY}}$$

$$= 271.60 - \frac{(198)^2}{274.40}$$

$$= (271.60) - (142.87)) = 128.73$$

$$\text{Adjusted } SS \text{ between groups} = (\text{Adjusted total } SS) - (\text{Adjusted } SS \text{ within group})$$

$$= (319.13 - 128.73)$$

$$= 190.40$$

ANOVA Table for Adjusted X

Source	d.f.	SS	MS	F-ratio
Between groups	2	190.40	95.2	8.14
Within group	11	128.73	11.7	
Total	13	319.13		

At 5% level, the table value of F for $v_1 = 2$ and $v_2 = 11$ is 3.98 and at 1% level the table value of s 7.21. Both these values are less than the calculated value (i.e., calculated value of 8.14 is greater an table values) and accordingly we infer that F-ratio is significant at both levels which means the fference in group means is significant.

Adjusted means on X will be worked out as follows:

$$\text{Regression coefficient for } X \text{ on } Y \text{ i.e., } b = \frac{\text{Sum of product within group}}{\text{Sum of squares within groups for } Y}$$

$$= \frac{198}{274.40} = 0.7216$$

Deviation of initial group means from general mean (= 14) in case of Y		Final means of groups in X (unadjusted)
Group I	–7.40	9.80
Group II	0.40	22.80
Group III	7.00	35.00

Adjusted means of groups in X = (Final mean) – b (Deviation of initial mean from general mean case of Y)

Hence,

Adjusted mean for Group I $= (9.80) - 0.7216\,(-7.4) = 15.14$

Adjusted mean for Group II $= (22.80) - 0.7216\,(0.40) = 22.51$

Adjusted mean for Group III $= (35.00) - 0.7216\,(7.00) = 29.95$

PROBLEMS

1. (a) Explain the meaning of analysis of variance. Describe briefly the technique of analysis of varian for one-way and two-way classifications.

   (b) State the basic assumptions of the analysis of variance.

2. What do you mean by the additive property of the technique of the analysis of variance? Explain ho this technique is superior in comparison to sampling.

3. Write short notes on the following:

   (i) Latin-square design.

   (ii) Coding in context of analysis of variance.

   (iii) *F*-ratio and its interpretation.

   (iv) Significance of the analysis of variance.

4. Below are given the yields per acre of wheat for six plots entering a crop competition, there of the plot being sown with wheat of variety *A* and three with *B*.

Variety	*Yields in fields per acre*		
	1	2	3
A	30	32	22
B	20	18	16

Set up a table of analysis of variance and calculate *F*. State whether the difference between the yields o two varieties is significant taking 7.71 as the table value of *F* at 5% level for $v_1 = 1$ and $v_2 = 4$.

5. A certain manure was used on four plots of land *A*, *B*, *C* and *D*. Four beds were prepared in each plot and the manure used. The output of the crop in the beds of plots *A*, *B*, *C* and *D* is given below:

Output on Plots

A	*B*	*C*	*D*
8	9	3	3
12	4	8	7
1	7	2	8
3	1	5	2

Find out whether the difference in the means of the production of crops of the plots is significant or not.

Present your conclusions after doing analysis of variance to the following results of the Latin-square design experiment conducted in respect of five fertilizers which were used on plots of different fertility.

A 16	*B* 10	*C* 11	*D* 09	*E* 09
E 10	*C* 09	*A* 14	*B* 12	*D* 11
B 15	*D* 08	*E* 08	*C* 10	*A* 18
D 12	*E* 06	*B* 13	*A* 13	*C* 12
C 13	*A* 11	*D* 10	*E* 07	*B* 14

Test the hypothesis at the 0.05 level of significance that $\mu_1 = \mu_2 = \mu_3$ for the following data:

Samples

No. one (1)	*No. two* (2)	*No. three* (3)
6	2	6
7	4	8
6	5	9
–	3	5
–	4	–
Total 19	18	28

8. Three varieties of wheat W_1, W_2 and W_3 are treated with four different fertilizers viz., f_1, f_2, f_3 and f_4. The yields of wheat per acre were as under:

Fertilizer treatment	*Varieties of wheat*			*Total*
	W_1	W_2	W_3	
f_1	55	72	47	174
f_2	64	66	53	183
f_3	58	57	74	189
f_4	59	57	58	174
Total	236	252	232	720

Set up a table for the analysis of variance and work out the F-ratios in respect of the above. Are the F-ratios significant?

9. The following table gives the monthly sales (in thousand rupees) of a certain firm in three states by four salesmen:

States	*Salesmen*				*Total*
	A	*B*	*C*	*D*	
X	5	4	4	7	20
Y	7	8	5	4	24
Z	9	6	6	7	28
Total	21	18	15	18	72

Set up an analysis of variance table for the above information. Calculate *F*-coefficients and sta whether the difference between sales affected by the four salesmen and difference between sal affected in three States are significant.

10. The following table illustrates the sample psychological health ratings of corporate executives in th field of Banking. Manufacturing and Fashion retailing:

Banking	41	53	54	55	43
Manufacturing	45	51	48	43	39
Fashion retailing	34	44	46	45	51

Can we consider the psychological health of corporate executives in the given three fields to be equa at 5% level of significance?

11. The following table shows the lives in hours of randomly selected electric lamps from four batches:

Batch	*Lives in hours*							
1	1600	1610	1650	1680	1700	1720	1800	
2	1580	1640	1640	1700	1750			
3	1450	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680		

Perform an analysis of variance of these data and show that a significance test does not reject thei homogeneity.

12. Is the interaction variation significant in case of the following information concerning mileage based o different brands of gasoline and cars?

		Brands of gasoline			
		W	*X*	*Y*	*Z*
Cars	*A*	13	12	12	11
		11	10	11	13
	B	12	10	11	9
		13	11	12	10
	C	14	11	13	10
		13	10	14	8

The following are paired observations for three experimental groups concerning an experimental involving three methods of teaching performed on a single class.

Method A to Group I		*Method B to Group II*		*Method C to Group III*	
X	*Y*	*X*	*Y*	*X*	*Y*
33	20	35	31	15	15
40	32	50	45	10	20
40	22	10	5	5	10
32	24	50	33	35	15

X represents initial measurement of achievement in a subject and *Y* the final measurement after subject has been taught. 12 pupils were assigned at random to 3 groups of 4 pupils each, one group from one method as shown in the table.

Apply the technique of analysis of covariance for analyzing the experimental results and then state whether the teaching methods differ significantly at 5% level. Also calculate the adjusted means on *Y*.

[**Ans:** *F*-ratio is not significant and hence there is no difference due to teaching methods.

Adjusted means on *Y* will be as under:

For Group I	20.70
For Group II	24.70
For Group III	22.60]

. The data in the table (in thousands of dollars) were extracted from Business Week's 1986 Executive Compensation Scoreboard. Assume that the data represent independent samples of 1986 total cash compensations for eight corporate executives in each of the three industries—banks, utilities, and office equipments/computers.

Banks	*Utilities*	*Office Equipments/Computers*
755	520	438
712	295	828
845	553	622
985	950	453
1300	930	562
1143	428	348
733	510	405
1189	864	938

Is there evidence of a difference among the means of 1986 total cash compensations for the three groups of corporate executives? Test at 1% level of significance. Find the industry (industries) for which mean compensation is (are) different from the others at 1% level.

13

Chapte

Other Nonparametric Methods

It has already been stated in earlier chapters that a statistical test is a formal technique, based some probability distributions, for arriving at a decision about the reasonableness of an assertio hypothesis. The test technique makes use of one or more values obtained from sample data [of called test statistic(s)] to arrive at a probability statement about the hypothesis. But such a technique also makes use of some more assertions about the population from which the sampl drawn. For instance, it may assume that population is normally distributed, sample drawn is a rand sample and similar other assumptions. The normality of the population distribution forms the basis making statistical inferences about the sample drawn from the population. But no such assumpti are made in case of non-parametric tests.

In a statistical test, two kinds of assertions are involved viz., an assertion directly related to purpose of investigation and other assertions to make a probability statement. The former is assertion to be tested and is technically called a hypothesis, whereas the set of all other assertion called the model. When we apply a test (to test the hypothesis) without a model, it is known distribution-free test, or the non-parametric test. Non-parametric tests do not make an assumpti about the parameters of the population and thus do not make use of the parameters of the distributi In other words, under non-parametric or distribution-free tests we do not assume that a particu distribution is applicable, or that a certain value is attached to a parameter of the population. instance, while testing the two training methods, say *A* and *B*, for determining the superiority of over the other, if we do not assume that the scores of the trainees are normally distributed or that mean score of all trainees taking method *A* would be a certain value, then the testing method known as a distribution-free or nonparametric method. In fact, there is a growing use of such tests situations when the normality assumption is open to doubt. As a result many distribution-free te have been developed that do not depend on the shape of the distribution or deal with the paramete of the underlying population. Some important nonparametric tests utilizing chi-square distributi have already been discussed in Chapter 11. Here we present some other nonparametric method

.1 SIGN TESTS

test is the nonparametric analogous of *t*-test where we test (i) if the population mean is the same pothetical mean or not (one sample case), and (ii) if the two population means are the same or two samples case). Here we do not give much weightage to the magnitude of the observations. central point of a data set is arithmetic mean when we give weightage to the magnitude of the rvation, while the location-wise central point is given by median. Therefore, in sign tests and r nonparametric tests we consider median in place of mean. We below explain one sample and sample sign tests.

1.1 One Sample Sign Test

ed on a single random sample coming from some continuous population, we wish to test Population median = M_0 against H_1: Population median > M_0, or H_1: Population median < M_0, or Population median ≠ M_0. Here, is M_0 the hypothetical value of the population mean. In the ple, each observation greater than M_0 is replaced by a '+' sign and each observation less than M_0 placed by a '–' sign. Observations which are equal to M_0 are ignored and sample size is reduced.

Let S = max. (No. of '+' signs, No. of '–' signs).

Under null hypothesis, the number of '+' signs and number of '–' signs are almost the same. refore under null hypothesis, S has Binomial (n, 1/2) distribution, where n is the size of (reduced) ple.

For one sided (left or right) tests, the p value is given by the probability of observing a value of S igher assuming Binomial (n, 1/2) distribution. For a two tailed test, the p value is twice of the bability obtained above. Binomial probabilities can be obtained using Table 9 given in the appendix e end of the book. For a given level of significance α, we reject null hypothesis when p value < α.

mple 13.1: Suppose playing four rounds of golf at a city club, 11 professionals totaled 280, 282, , 273, 283, 283, 275, 284, 282, 279, and 281. At 5% level of significance, test that the professional ers average 284 for four rounds against they average less.

ution: We shall use nonparametric method for this as we do not know if the parent population is mal or not. Sample is so small that by making a histogram we may not get a reasonable idea of ent population. In nonparametric tests, the median is used for average.

We wish to test H_0: Population median = 284 against H_1: Population median < 284.

In nonparametric methods, we do not obtain the estimates of parameters as in case of parametric s. Here, we do not calculate sample median.

The observations greater than 284 are replaced by a '+' signs and the observations less than 284 replaced by '–' signs, and the observations equal to 284 are discarded. Thus we have a sequence igns as below:

$$-,\ -,\ +,\ -,\ -,\ -,\ -,\ -,\ -,\ -$$

One observation equals to 283, is discarded and resulting sample size is n = 10 instead of 11.

S = max (No. of '+' signs, No. of '–' signs) = 9.

p-value for this one sided test.

= Prob [Observing a value of 9 or higher using B(10, 1/2)]

= 1 – Prob [Observing a value of 8 or less using B(10, 1/2)]

$= 1 - 0.9892 = 0.0108.$

Since, p value < 0.05, we reject the null hypothesis at 5% level of significance and conclude the golfers' average is less than 284 in four rounds.

Note: In case of large sample size (more than 30), binomial probabilities can be approxima using normal distribution.

Probability of observing a value of S or higher assuming Binomial (n, 1/2) distribution

$$= P\left(Z \geq \frac{s - \frac{n+1}{2}}{\sqrt{n/4}}\right) = 1 - P\left(Z < \frac{s - \frac{n+1}{2}}{\sqrt{n/4}}\right),$$ where Z has N(0, 1) distribution.

13.1.2 Two Samples Sign Test (Paired Sign Test)

This test is an analogous of paired t-test. We use this test when we have two samples having pa observations. The usual null hypothesis for this test is that there is no difference between the locat of the centers of two populations, i.e., H_0: $M_1 = M_2$ against H_1: $M_1 > M_2$, or H_1: $M_1 < M_2$, or H_1: $\neq M_2$. Here, are the notations M_1 and M_2 for population medians.

For each pair, subtract the 2nd score from the 1st, and write down the sign of the differe Discard the pairs having both observations as the same. Follow the procedure of Sign test for si population for hypothetical median as zero. The method is illustrated using an example below:

Example 13.2: The following are the numbers of aircrafts dug up by two archaeologists. A and at an ancient cliff dwelling on 30 days.

A	1	0	2	3	1	0	2	2	3	0	1	1	4	1	2	1	3	5	2	1	3	2	4	1	3	2	0	2	4
B	0	0	1	0	2	0	0	1	1	2	0	1	2	1	1	0	2	2	6	0	2	3	0	2	1	0	1	0	1

Use sign test at 1% level of significance to test that the two archaeologists. A and B. are equa good in finding artifacts against A is better.

Solution: We wish to test H_0: $M_1 = M_2$ against H_1: $M_1 > M_2$.

We first make the following table:

Table 13.1

A	*B*	*Sign of (A − B)*
1	0	+
0	0	0
2	1	+
3	0	+
1	2	−
0	0	0
2	0	+
2	1	+
3	1	+
0	2	−
1	0	+
4	2	+
1	1	0
2	1	+
1	0	+
3	2	+
5	2	+
2	6	−
1	0	+
3	2	+
2	3	−
4	0	+
1	2	−
3	1	+
2	0	+
0	1	−
2	0	+
4	1	+
2	0	+

Number of '+' signs = 20,

Number of '−' signs = 6,

Number of 0 = 4,

Reduced sample size $n = 26$.

S = max. (20, 6) = 20.

p-value for this one sided test

= Prob [Observing a value of 20 or higher using B(26, 1/2)]

= 1 − Prob [Observing a value of 19 or less using B(26, 1/2)]

= 1 − 0.995322 = 0.004678.

Using normal approximation, this probability is obtained as

$$P\left(Z \geq \frac{s-\frac{n+1}{2}}{\sqrt{n/4}}\right) = 1 - P\left(Z < \frac{s-\frac{n+1}{2}}{\sqrt{n/4}}\right)$$

$$= 1 - P(Z < 2.54951) = 1 - 0.994606 = 0.005394$$

Since, p-value < 0.05, we reject the null hypothesis at 5% level of significance and conclude t artifact *A* is better.

13.2 WILCOXON SIGNED RANK SUM TEST FOR SINGLE POPULATION

Sign test does not give any weightage to the magnitude of the observation. So, the results of Sign t are much weaker. The Wilcoxon signed rank sum test is used in place of one-sample *t*-test a paired *t*-test. This test can also be used for ordered categorical data where a numerical scale inappropriate but it is possible to rank the observations.

13.2.1 Wilcoxon Signed Rank Sum Test (Single Sample)

In Wilcoxon signed rank test (single sample), we test the same hypothesis, i.e., H_0: Population medi $= M_0$ against H_1: Population median $> M_0$, or H_1: Population median $< M_0$, or H_1: Population medi $\neq M_0$. The test is described as below:

(i) We first obtain the difference of each observation from hypothetical M_0. If some differenc are zero, we discard them and reduce the sample size accordingly. Suppose, *n* is the si of (reduced) sample.

(ii) Now we assign the ranks 1, 2,..., *n* to the absolute differences (ignoring the sign) fro smallest to largest. Do not lose the track of the signs of differences.

(iii) If you have ties in ranks, assign the average value of those ranks that would be assigned case of no ties.

(iv) Calculate W^+ = Sum of the ranks corresponding to positive differences and W^- = Sum the ranks corresponding to negative differences. (As a check the total, $W^+ + W^-$ = Shou be equal to $n(n + 1)/2$).

It is expected under the null hypothesis that the distribution of the differences is almo symmetric around zero and the distribution of positives and negatives to be distributed random among the ranks. We use this concept to make the decision.

(v) We reject null hypothesis at $\alpha \times 100\%$ level of significance if

$W^- \leq W_\alpha$, for a right tailed test;

$W^+ \leq W_\alpha$ for a left tailed test;

$W^- \leq W_{\alpha/2}$ or $W^+ \leq W_{\alpha/2}$, for a two tailed test;

The value of W_α and $W_{\alpha/2}$ are given in Table in the appendix given at the end of the book.

Note: In case of large samples, the distribution of $\dfrac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$ is N(0, 1), here W = min (W^-, W^+).

ample 13.3: Solve Example 13.1 using Wilcoxon test.

ution: We have the observations 280, 282, 292, 273, 283, 283, 275, 284, 282, 279, and 281.

We wish to test H_0: Population median = 284 against H_1: Population median < 284.

We first obtain the difference of each observation 284 and assign the ranks as below:

Table 13.2

S. No.	*Observation*	*Difference = Observations − 284*	*Absolute difference*	*Rank*
1	280	− 4	4	6
2	**282**	**− 2**	**2**	**3.5**
3	292	8	8	8
4	273	−11	11	10
5	**283**	**−1**	**1**	**1.5**
6	**283**	**−1**	**1**	**1.5**
7	275	−9	9	9
8	284	0	0	–
9	**282**	**−2**	**2**	**3.5**
10	279	−5	5	7
11	281	−3	3	5

Since one difference is zero, so we discard the corresponding observation from the sample. duced sample size is $n = 10$. We have tied ranks at two places as absolute difference 2 is repeated ice and absolute difference 1 is repeated twice. As discussed we assign average ranks for tied solute differences.

We calculate $W^+ = 8$ and $W^- = 47$. (As a check, we verify that $W^+ + W^- = n(n + 1)/2 = 55$.)

At 5% level of significance, the critical value from Table, $W_{0.05} = 11$. Since $W^+ \le W_\alpha$, we reject e null hypothesis and conclude that the golfers' average is less than 284 in four rounds.

.2.2 Wilcoxon Signed Rank Sum Test (Paired Samples) or Wilcoxon Matched Pairs Test

is test is analogous to paired t-test and stronger than the sign test for paired observation. We wish test H_0: $M_1 = M_2$ against H_1: $M_1 > M_2$, or H_1: $M_1 < M_2$, or H_1: $M_1 \neq M_2$. Here, are the notations M_1 d M_2 for population medians.

For each pair, obtain the difference by subtracting the 2nd score from the 1st. Use the obtai differences and follow the procedure of Wilcoxon signed rank sum test for one sample us hypothetical median as zero. The method is illustrated using an example below:

Example 13.4: An experiment is conducted to judge the effect of brand name on quality percepti 16 subjects are recruited for the purpose and are asked to taste and compare two samples of prod on a set of scale items judged to be ordinal. The following data are obtained:

Pair	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1
Brand A	73	43	47	53	58	47	52	58	38	61	56	56	34	55	65	7
Brand B	51	41	43	41	47	32	24	58	43	53	52	57	44	57	40	6

Test the hypothesis, using Wilcoxon matched pairs test, that there is no difference between perceived qualities of the two brands. Use 5% level of significance.

Solution: We wish to test

H_0: no difference between the perceived qualities of the two brands

Against

H_1: significant difference between the perceived qualities of the two brands.

We obtain the differences by subtracting the scores of Brand B from the corresponding score Brand A and assign the ranks to the absolute differences as below:

Table 13.3

Pair	Brand A	Brand B	Difference	Rank
1	73	51	22	13
2	43	41	2	2.5
3	47	43	4	4.5
4	53	41	12	11
5	58	47	11	10
6	47	32	15	12
7	52	24	28	15
8	58	58	0	–
9	38	43	–5	6
10	61	53	8	8
11	56	52	4	4.5
12	56	57	–1	1
13	34	44	–10	9
14	55	57	–2	2.5
15	65	40	25	14
16	75	68	7	7

We calculate $W^+ = 101.5$ and $W^- = 18.5$. (As a check, we verify that $W^+ + W^- = n(n+1)/2 =$ 0.) Also, the reduced sample size $n = 15$. For this two tailed test at 5% level of significance, the tical value from Table is 26 which is larger than W^- therefore we reject the null hypothesis.

3.3 MANN WHITNEY U TEST

is test is an analogous of t-test for two independent samples. Here we test H_0: $M_1 = M_2$ against : $M_1 > M_2$, or H_1: $M_1 < M_2$, or H_1: $M_1 \neq M_2$ on the basis of two independent samples drawn from ntinuous populations.

To perform this test, we first of all rank the data jointly, taking them as belonging to a single mple in either an increasing or decreasing order of magnitude. We usually adopt low to high nking process which means we assign rank 1 to an item with lowest value, rank 2 to the next higher m and so on. In case there are ties, then we would assign each of the tied observation the mean of e ranks which they jointly occupy. For example, if sixth, seventh and eighth values are identical, we ould assign each the rank $(6 + 7 + 8)/3 = 7$. After this we find the sum of the ranks assigned to the lues of the first sample (and call it R_1) and also the sum of the ranks assigned to the values of the cond sample (and call it R_2). Then we work out the test statistic i.e., U, which is a measurement of e difference between the ranked observations of the two samples as under:

$$U = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

here n_1, and n_2 are the sample sizes and R_1 is the sum of ranks assigned to the values of the first mple. (In practice, whichever rank sum can be conveniently obtained can be taken as R_1, since it immaterial which sample is called the first sample.)

In applying U-test we take the null hypothesis that the two samples come from identical populations. this hypothesis is true, it seems reasonable to suppose that the means of the ranks assigned to the lues of the two samples should be more or less the same. Under the alternative hypothesis, the eans of the two populations are not equal and if this is so, then most of the smaller ranks will go to e values of one sample while most of the higher ranks will go to those of the other sample.

If the null hypothesis that the $n_1 + n_2$ observations came from identical populations is true, the id 'U' statistic has a sampling distribution with

$$\text{Mean} = \mu_U = \frac{n_1 \cdot n_2}{2}$$

nd Standard deviation (or the standard error)

$$= \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

n_1 and n_2 are sufficiently large (i.e., both greater than 8), the sampling distribution of U can be pproximated closely with normal distribution and the limits of the acceptance region can be determined the usual way at a given level of significance. But if either n_1 or n_2 is so small that the normal curve pproximation to the sampling distribution of U cannot be used. In such a case, we use the exact distribution based on the values Table 6 given in the appendix at the end of the book. How to ap the test in small samples is explained in Example 13.6. We present first a large sample exam below:

Example 13.5: The values in one sample are 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. another sample they are 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. Test at the 10% level hypothesis that they come from populations with the same mean. Apply *U*-test.

Solution: First of all we assign ranks to all observations, adopting low to high ranking process on t presumption that all given items belong to a single sample. By doing so we get the following:

Table 13.4

Size of sample item in ascending order	*Rank*	*Name of related sample: [A for sample one and B for sample two]*
32	1	*B*
38	2	*A*
39	3	*A*
40	4	*B*
41	5	*B*
44	6.5	*B*
44	6.5	*B*
46	8	*A*
48	9	*A*
52	10	*B*
53	11.5	*B*
53	11.5	*A*
57	13	*A*
60	14	*A*
61	15	*B*
67	16	*B*
69	17	*A*
70	18	*B*
72	19.5	*B*
72	19.5	*B*
73	21.5	*A*
73	21.5	*A*
74	23	*A*
78	24	*A*

From the above we find that the sum of the ranks assigned to sample one items or $R_1 = 2 + 3 +$ $+ 11.5 + 13 + 14 + 17 + 21.5 + 21.5 + 23 + 24 = 167.5$ and similarly we find that the sum of ranks gned to sample two items or $R_2 = 1 + 4 + 5 + 6.5 + 6.5 + 10 + 11.5 + 15 + 16 + 18 + 19.5 + 19.5$ 2.5 and we have $n_1 = 12$ and $n_2 = 12$.

Hence, test statistic $U = n_1 \cdot n_2 + \dfrac{n_1(n_1+1)}{2} - R_1$

$$= (12)\,(12) + \frac{12(12+1)}{2} - 167.5$$

$$= 144 + 78 - 167.5 = 54.5$$

Since in the given problem n_1 and n_2 both are greater than 8, so the sampling distribution of U roximates closely with normal curve. Keeping this in view, we work out the mean and standard iation taking the null hypothesis that the two samples come from identical populations as under:

$$\mu_U = \frac{n_1 \times n_2}{2} = \frac{(12)\,(12)}{2} = 72$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(12)\,(12)\,(12+12+1)}{12}}$$

$$= 17.32$$

As the alternative hypothesis is that the means of the two populations are not equal, a two-tailed t is appropriate.

Now we calculate the test statistic $Z_c = \dfrac{U - \mu_U}{\sigma_U} = \dfrac{54.5 - 72}{17.32} = -1.0104$

e p-value for this two tailed test using N(0,1) distribution is $2P(Z > 1.0104) = 0.3123$ which is ger than 0.1. Therefore, at 10% level of significance, we can not reject the null hypothesis.

ample 13.6: Two samples with values 90, 94, 36 and 44 in one case and the other with values 53, 6, 24, and 33 are given. Test applying Wilcoxon test whether the two samples come from populations th the same mean at 10% level against the alternative hypothesis that these samples come from pulations with different medians.

lution: Let us first assign ranks as stated earlier and we get:

Table 13.5

Size of sample item in ascending order	*Rank*	*Name of related sampl (Sample one as A Sample two as B)*
6	1	*B*
24	2	*B*
33	3	*B*
36	4	*A*
39	5	*B*
44	6	*A*
53	7	*B*
90	8	*A*
94	9	*A*

Sum of ranks assigned to items of sample one = 4 + 6 + 8 + 9 = 27
Number of items in this sample = 4
Sum of ranks assigned to items of sample two = 1 + 2 + 3 + 5 + 7 = 18
Number of items in this sample = 5

As the number of items in the two samples is less than 8, we cannot use the normal cu approximation technique as stated above and shall use the table giving values of Wilcoxon's distribut To use this table, we denote 'W_s' as the smaller of the two sums and 'W_l' the larger. Also, let 's the number of items in the sample with smaller sum and let '*l*' be the number of items in the sam with the larger sum. Taking these notations we have for our question the following values:

$$W_s = 18;\ s = 5;\ W_l = 27;\ l = 4$$

The value of W_s is 18 for sample two which has five items and as such $s = 5$. We now find difference between W_s and the minimum value it might have taken, given the value of *s*. The minim value that W_s could have taken, given that $s = 5$, is the sum of ranks 1 through 5 and this come equal to $1 + 2 + 3 + 4 + 5 = 15$. Thus, (W_s – Minimum W_s) = 18 – 15 = 3. To determine the probabi that a result as extreme as this or more so would occur, we find the cell of the table which is in column headed by the number 3 and in the row for $s = 5$ and $l = 4$ (the specified values of *l* are gi in the second column of the table). The entry in this cell is 0.056 which is the required probabilit getting a value as small as or smaller than 3 and now we should compare it with the significance le of 10%. Since the alternative hypothesis is that the two samples come from populations with differ means, a two-tailed test is appropriate and accordingly 10% significance level will mean 5% in left tail and 5% in the right tail. In other words, we should compare the calculated probability with probability of 0.05, given the null hypothesis and the significance level. If the calculated probabil happens to be greater than 0.05 (which actually is so in the given case as 0.056 > 0.05), then should accept the null hypothesis. Hence, in the given problem, we must conclude that the t samples come from populations with the same mean.

(The same result we can get by using the value of W_l. The only difference is that the va maximum W_l – W_l is required. Since for this problem, the maximum value of W_l (given $s = 5$ a $l = 4$) is the sum of 6 through 9 i.e., 6 + 7 + 8 + 9 = 30, we have Max. W_l – W_l = 30 – 27 = 3 wh is the same value that we worked out earlier as W_s – Minimum W_s. All other things then remain same as we have stated above).

.4 RUN TEST

s test is a test used to judge the randomness of a sample on the basis of the order in which the ervations are taken. There are many applications in which it is difficult to decide whether the ple used is a random one or not. This is particularly true when we have little or no control over selection of the data. For instance, if we want to predict a retail store's sales volume for a given nth, we have no choice but to use past sales data and perhaps prevailing conditions in general. ne of this information constitutes a random sample in the strict sense. To allow us to test samples the randomness of their order, statisticians have developed the theory of runs. A run is a succession dentical letters (or other kinds of symbols) which is followed and preceded by different letters or letters at all. To illustrate, we take the following arrangement of healthy, *H*, and diseased, *D*, ngo trees that were planted many years ago along a certain road:

$$\underset{\text{1st}}{\underline{HH}}\quad \underset{\text{2nd}}{\underline{DD}}\quad \underset{\text{3rd}}{\underline{HHHHH}}\quad \underset{\text{4th}}{\underline{DDD}}\quad \underset{\text{5th}}{\underline{HHHH}}\quad \underset{\text{6th}}{\underline{DDDDD}}\quad \underset{\text{7th}}{\underline{HHHHHHHHH}}$$

ing underlines to combine the letters which constitute the runs, we find that first there is a run of o *H*'s, then a run of two *D*'s, then a run of five *H*'s, then a run of three *D*'s, then a run of four *H*'s, en a run of five *D*'s and finally a run of nine *H*'s. In this way there are 7 runs in all or $r = 7$. If there too few runs, we might suspect a definite grouping or a trend; if there are too many runs, we ght suspect some sort of repeated alternating patterns. In the given case there seems some grouping ., the diseased trees seem to come in groups. Through one sample runs test which is based on the ea that too few or too many runs show that the items were not chosen randomly, we can say nether the apparently seen grouping is significant or whether it can be attributed to chance. We all use the following symbols for a test of runs:

n_1 = Number of occurrences of type 1 (say *H* in the given case)
n_2 = Number of occurrences of type 2 (say *D* in the given case)
r = Number of runs.

In the given case the values of n_1, n_2 and r would be as follows:

$$n_1 = 20;\ n_2 = 10;\ r = 7$$

The sampling distribution of '*r*' statistic, the number of runs, is to be used and this distribution has s mean

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$

nd the standard deviation $\sigma_r = \sqrt{2n_1n_2\dfrac{(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}}$

In the given case, we work out the values of μ_r and σ_r as follows:

$$\mu_r = \frac{(2)\,(20)\,(10)}{20 + 10} + 1 = 14.33$$

nd

$$\sigma_r = \sqrt{\frac{(2)\,(20)\,(10)\,(2 \times 20 \times 10 - 20 -- 10)}{(20 + 10)^2\,(20 + 10 - 1)}} = 2.38$$

For testing the null hypothesis concerning the randomness of the planted trees, we should h been given the level of significance. Suppose it is 1% or 0.01. Since too many or too few runs wo indicate that the process by which the trees were planted was not random, a two-tailed tes appropriate which can be indicated as follows on the assumption that the sampling distribution can be closely approximated by the normal distribution. This assumption can be applied when n_1 n_2 are sufficiently large i.e., they should not be less than 10. But in case n_1 or n_2 is so small that normal curve approximation assumption cannot be used, then exact tests may be based on spe tables which can be seen in the book *Non-parametric Statistics for the Behavioural Science* by Siegel.

$$\text{Now we will calculate } Z_c = \frac{r-\mu_r}{\sigma_r} = \frac{7-14.33}{2.38} = -3.0798$$

The *p*-value for this two tailed test using N(0, 1) distribution is $2P(Z > 3.0798) = 0.0021$ wh is much smaller than 0.01. Therefore, at 1% level of significance, we can reject the null hypothe We conclude that there is a strong indication that the diseased trees come in a non-random groupi

One sample runs test, as explained above, is not limited only to test the randomness of series attributes. Even a sample consisting of numerical values can be treated similarly by using the lett say '*a*' and '*b*' to denote respectively the values falling above and below the median of the samp Numbers equal to the median are omitted. The resulting series of *a*'s and *b*'s (representing the da in their original order) can be tested for randomness on the basis of the total number of runs abo and below the median, as per the procedure explained above.

(The method of runs above and below the median is helpful in testing for trends or cyclic patterns concerning economic data. In case of an upward trend, there will be first mostly *b*'s a later mostly *a*'s, but in case of a downward trend, there will be first mostly *a*'s and later mostly *b* In case of a cyclical pattern, there will be a systematic alternating of *a*'s and *b*'s and probably ma runs.)

13.5 KRUSKAL WALLIS TEST

This test is conducted in a way similar to the *U* test described above. This test is used to test the nu hypothesis that '*k*' independent random samples come from identical universes against the alternativ hypothesis that the medians of these universes are not equal. This test is analogous to the one-wa analysis of variance, but unlike the latter it does not require the assumption that the samples com from approximately normal populations or the universes having the same standard deviation.

In this test, like the *U* test, the data are ranked jointly from low to high or high to low as if the constituted a single sample. The test statistic is *H* for this test which is worked out as under:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where, $n = n_1 + n_2 + \ldots + n_k$ and R_i being the sum of the ranks assigned to n_i observations in the it sample.

If the null hypothesis is true that there is no difference between sample medians and each ple has at least five items, then the sampling distribution of H can be approximated with a chi- are distribution with $(k - 1)$ degrees of freedom. As such we can reject the null hypothesis at a en level of significance if H value calculated, as stated above, exceeds the concerned table value hi-square.For small sample the critical values can be taken from the Table given in the appendix ne end of this book. Let us take an example to explain the operation of this test.

ample 13.7 : Use the Kruskal-Wallis test at 5% level of significance to test the null hypothesis that ofessional bowler performs equally well with the four bowling balls, given the following results:

Bowling Results in Five Games

With Ball No. *A*	271	282	257	248	262
With Ball No. *B*	252	275	302	268	276
With Ball No. *C*	260	255	239	246	266
With Ball No. *D*	279	242	297	270	258

ution: To apply the H test or the Kruskal-Wallis test to this problem, we begin by ranking all the en figures from the highest to the lowest, indicating besides each the name of the ball as under:

Table 13.6

Bowling results	*Rank*	*Name of the ball associated*
302	1	*B*
297	2	*D*
282	3	*A*
279	4	*D*
276	5	*B*
275	6	*B*
271	7	*A*
270	8	*D*
268	9	*B*
266	10	*C*
262	11	*A*
260	12	*C*
258	13	*D*
257	14	*A*
255	15	*C*
252	16	*B*
248	17	*A*
246	18	*C*
242	19	*D*
239	20	*C*

For finding the values of R_i, we arrange the above table as under:

Table 13.7: Bowling Results with Different Balls and Corresponding Rank

Ball A	Rank	Ball B	Rank	Ball C	Rank	Ball D	Rank
271	7	252	16	260	12	279	4
282	3	275	6	255	15	242	19
257	14	302	1	239	20	297	2
248	17	268	9	246	18	270	8
262	11	276	5	266	10	158	13
$n_1 = 5$	$R_1 = 52$	$n_2 = 5$	$R_2 = 37$	$n_3 = 5$	$R_3 = 75$	$n_4 = 5$	$R_4 = 46$

Now we calculate H statistic as under:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{20(20+1)} \left\{ \frac{52^2}{5} + \frac{37^2}{5} + \frac{75^2}{5} + \frac{46^2}{5} \right\} - 3(20+1)$$

$$= (0.02857)\,(2362.8) - 63 = 67.51 - 63 = 4.51$$

As the four samples have five items each, the sampling distribution of H approximates clos with χ^2 distribution. Now taking the null hypothesis that the bowler performs equally well with four balls, we have the value of $\chi^2 = 7.815$ for $(k - 1)$ or $4 - 1 = 3$ degrees of freedom at 5% le of significance. Since the calculated value of H is only 4.51 and does not exceed the χ^2 value 7.815, so we accept the null hypothesis and conclude that bowler performs equally well with the f bowling balls.

13.6 SPEARMAN'S RANK CORRELATION

When the data are not available to use in numerical form for doing correlation analysis but when t information is sufficient to rank the data as first, second, third, and so forth, we quite often use t rank correlation method and work out the coefficient of rank correlation. In fact, the rank correlati coefficient is a measure of correlation that exists between the two sets of ranks. In other words, it a measure of association that is based on the ranks of the observations and not on the numeric values of the data. It was developed by famous statistician Charles Spearman in the early 1900s a as such it is also known as Spearman's rank correlation coefficient.

For calculating rank correlation coefficient, first of all the actual observations be replaced their ranks, giving rank 1 to the highest value, rank 2 to the next highest value and following this ve order ranks are assigned for all values. If two or more values happen to be equal, then the avera of the ranks which should have been assigned to such values had they been all different, is taken a same rank (equal to the said average) is given to concerning values. The second step is to record difference between ranks (or 'd') for each pair of observations, then square these differences to ain a total of such differences which can symbolically be stated as $\sum d_i^2$. Finally, Spearman's rank relation coefficient, r^*, is worked out as under:

$$\text{Spearman's } `r' = 1 - \left\{\frac{6\sum d_i^2}{n\left(n^2 - 1\right)}\right\}$$

ere, n = Number of paired observations.

The value of Spearman's rank correlation coefficient will always vary between ± 1, +1, indicating erfect positive correlation and –1 indicating perfect negative correlation between two variables. l other values of correlation coefficient will show different degrees of correlation.

Suppose we get $r = 0.756$ which suggests a substantial positive relationship between the concerning o variables. But how we should test this value of 0.756? The testing device depends upon the lue of n. For small values of n (i.e., n less than 30), the distribution of r is not normal and as such use the table showing the values for Spearman's Rank correlation (Table No. 5 given in Appendix the end of the book) to determine the acceptance and rejection regions. Suppose we get $r = 0.756$ r a problem where $n = 15$ and want to test at 5% level of significance the null hypothesis that there zero correlation in the concerning ranked data. In this case our problem is reduced to test the null pothesis that there is no correlation i.e., $u_r = 0$ against the alternative hypothesis that there is a rrelation i.e., $\mu_r \neq 0$ at 5% level. In this case a two-tailed test is appropriate and we look in the id table in row for $n = 15$ and the column for a significance level of 0.05 and find that the critical lues for r are ± 0.5179 i.e., the upper limit of the acceptance region is 0.5179 and the lower limit the acceptance region is –0.5179. And since our calculated $r = 0.756$ is outside the limits of the ceptance region, we reject the null hypothesis and accept the alternative hypothesis that there is a rrelation in the ranked data.

In case the sample consists of more than 30 items, then the sampling distribution of r is pproximately normal with a mean of zero and a standard deviation of $1/\sqrt{n-1}$ and thus, the andard error of r is:

$$\sigma_r = \frac{1}{\sqrt{n-1}}$$

---

[*] Some authors use the symbol Rho (ρ) for this coefficient. Rho is to be used when the sample size does not exceed 30.

We can use the table of area under normal curve to find the appropriate z values for test hypotheses about the population rank correlation and draw inference as usual. We can illustrate it an example.

Example 13.8: Personnel manager of a certain company wants to hire 30 additional programm for his corporation. In the past, hiring decisions had been made on the basis of interview and also the basis of an aptitude test. The agency doing aptitude test had charged Rs. 100 for each test, b now wants Rs. 200 for a test. Performance on the test has been a good predictor of a programme ability and Rs. 100 for a test was a reasonable price. But now the personnel manager is not sure t the test results are worth Rs. 200. However, he has kept over the past few years records of scores assigned to applicants for programming positions on the basis of interviews taken by him he becomes confident (using 0.01 level of significance) that the rank correlation between his intervi scores and the applicants' scores on aptitude test is positive, then he will feel justified in discontinui the aptitude test in view of the increased cost of the test. What decision should he take on the ba of the following sample data concerning 35 applicants?

Sample Data Concerning 35 Applicants

Serial Number	*Interview score*	*Aptitude test score*
1	81	113
2	88	88
3	55	76
4	83	129
5	78	99
6	93	142
7	65	93
8	87	136
9	95	82
10	76	91
11	60	83
12	85	96
13	93	126
14	66	108
15	90	95
16	69	65
17	87	96
18	68	101
19	81	111
20	84	121
21	82	83
22	90	79
23	63	71
24	78	109

Serial Number	*Interview score*	*Aptitude test score*
25	73	68
26	79	121
27	72	109
28	95	121
29	81	140
30	87	132
31	93	135
32	85	143
33	91	118
34	94	147
35	94	138

lution: To solve this problem we should first work out the value of Spearman's *r* as under:

Table 13.8: Calculation of Spearman's

S. No.	*Interview score X*	*Aptitude test score Y*	*Rank X*	*Rank Y*	*Rank difference 'd$_i$' (Rank X) – (Rank Y)*	*Differences squared d_i^2*
1	81	113	21	15	6	36
2	88	88	11	27	–16	256
3	55	76	35	32	3	9
4	83	129	18	9	9	81
5	78	99	24.5	21	3.5	12.25
6	93	142	6	3	3	9
7	65	93	32	25	7	49
8	87	136	13	6	7	49
9	95	82	1.5	30	–28.5	812.25
10	76	91	26	26	0	0
11	60	83	34	28.5	5.5	30.25
12	85	96	15.5	22.5	–7	49
13	93	126	6	10	–4	16
14	66	108	31	18.5	12.5	156.25
15	90	95	9.5	24	–14.5	210.25
16	69	65	29	35	–6	36
17	87	96	13	22.5	–9.5	90.25
18	68	101	30	20	10	100
19	81	111	21	16	5	25
20	84	121	17	12	5	25
21	82	83	19	28.5	–9.5	90.25

S. No.	Interview score X	Aptitude test score Y	Rank X	Rank Y	Rank difference 'd' (Rank X) – (Rank Y)	Differences squared d^2
22	90	79	9.5	31	−21.5	462.25
23	63	71	33	33	0	0
24	78	108	24.5	18.5	6	36
25	73	68	27	34	−7	49
26	79	121	23	12	11	121
27	72	109	28	17	11	121
28	95	121	1.5	12	−10.5	110.25
29	81	140	21	4	17	289
30	87	132	13	8	5	25
31	93	135	6	7	−1	1
32	85	143	15.5	2	13.5	182.25
33	91	118	8	14	−6	36
34	94	147	3.5	1	2.5	6.25
35	94	138	3.5	5	−1.5	225
$n = 35$						$\Sigma d_i^2 = 3583$

$$\text{Spearman's 'r'} = 1 - \left\{\frac{6\Sigma d_i^2}{n\left(n^2 - 1\right)}\right\} = 1 - \left\{\frac{6 \times 3583}{35\left(35^2 - 1\right)}\right\}$$

$$= 1 - \frac{21498}{42840} = 0.498$$

Since $n = 35$ the sampling distribution of r is approximately normal with a mean of zero and standard deviation of $1/\sqrt{n-1}$. Hence the standard error of r is

$$\sigma_r = \frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{35-1}} = 0.1715$$

As the personnel manager wishes to test his hypothesis at 0.01 level of significance, the problem can be stated:

Null hypothesis that there is no correlation between interview score and aptitude test score i.e. $\mu_r = 0$.

Alternative hypothesis that there is positive correlation between interview score and aptitude test score i.e., $\mu_r > 0$.

Now we will calculate $Z_c = \dfrac{r - \mu_r}{\sigma_r} = \dfrac{0.498 - 0}{0.1715} = 2.9038$

The p-value for this right tailed test using $N(0, 1)$ distribution is $P(Z > 2.9038) = 0.0018$ which is h smaller than 0.01. Therefore, at 1% level of significance, we can reject the null hypothesis. conclude that the correlation between interview score and aptitude test score is positive. ordingly, personal manager may decide on discontinuing the aptitude test.

.7 SOME FEATURES OF NON-PARAMETRIC TESTS

m what has been stated above in respect of important non-parametric tests, we can say that se tests share in main the following characteristics:

1. They do not suppose any particular distribution and the consequential assumptions.
2. They are rather quick and easy to use i.e., they do not require laborious computations since in many cases the observations are replaced by their rank order and in many others we simply use signs.
3. They are often not as efficient or 'sharp' as tests of significance or the parametric tests. An interval estimate with 95% confidence may be twice as large with the use of non-parametric tests as with regular standard methods. The reason being that these tests do not use all the available information but rather use groupings or rankings and the price we pay is a loss in efficiency. In fact, when we use non-parametric tests, we make a trade-off: we loose sharpness in estimating intervals, but we gain the ability to use less information and to calculate faster.
4. When our measurements are not as accurate as is necessary for standard tests of significance, then non-parametric methods come to our rescue which can be used fairly satisfactorily.
5. Parametric tests cannot apply to ordinal or nominal scale data but non-parametric tests do not suffer from any such limitation.
6. The parametric tests of difference like 't' or 'F' make assumption about the homogeneity of the variances whereas this is not necessary for non-parametric tests of difference.

There are many situations in which the various assumptions required for standard tests of gnificance (such as that population is normal, samples are independent, standard deviation is known, tc.) cannot be met, then we can use non-parametric methods. Moreover, they are easier to explain nd easier to understand. This is the reason why such tests have become popular. But one should not orget the fact that they are usually less efficient/powerful as they are based on no assumption (or irtually no assumption) and we all know that the less one assumes, the less one can infer from a set f data. But then the other side must also be kept in view that the more one assumes, the more one imits the applicability of one's methods.

PROBLEMS

1. Give your understanding of non-parametric or distribution-free methods explaining their import characteristics.
2. Narrate the various advantages of using non-parametric tests. Also point out their limitations.
3. Briefly describe the different non-parametric tests explaining the significance of each such test.
4. On 15 occasions Mr. Kalicharan had to wait 4, 8, 2, 7, 7, 5, 8, 6, 1, 9, 6, 6, 5, 9 and 5 minutes for the bus takes to reach his office. Use the sign test at 5% level of significance to test the bus company's cla that on the average Mr. Kalicharan should not have to wait more than 5 minutes for a bus.
5. The following are the numbers of tickets issued by two policemen in 20 days:

By first policeman	7, 10, 14, 12, 6, 9, 11, 13, 7, 6, 10, 8, 14, 8, 12, 11, 9, 8, 10 and 15.
By second policeman	10, 13, 14, 11, 10, 7, 15, 11, 10, 9, 8, 12, 16, 10, 10, 14, 10, 12, 8 and 14.

Use the sign test at 1% level of significance to test the null hypothesis that on the average the tw policemen issue equal number of tickets against the alternative hypothesis that on the average th second policeman issues more tickets than the first one.

6. The following are the kilometres per gallon which a test driver got for ten tankfuls each of three kinds c gasoline:

Gasoline *A*	30,	41,	34,	43,	33,	34,	38,	26,	29,	36
Gasoline *B*	39,	28,	39,	29,	30,	31,	44,	43,	40,	33
Gasoline *C*	29,	41,	26,	36,	41,	43,	38,	38,	35,	40.

Use the Kruskal-Wallis test at the level of significance $\alpha = 0.05$ to test the null hypothesis that there i no difference in the average kilometre yield of the three types of gasoline.

7. (a) The following are the number of students absent from a college on 24 consecutive days: 29, 25, 31, 28, 30, 28, 33, 31, 35, 29, 31, 33, 35, 28, 36, 30, 33, 26, 30, 28, 32, 31, 38 and 27. Test for randomness at 1% level of significance.

   (b) The following arrangement indicates whether 25 consecutive persons interviewed by a social scientist are for (*F*) or against (*A*) an increase in the number of crimes in a certain locality: *F, F, F, F, F, F, A, F, F, F, F, F, A, F, F, F, F, A, A, F, F, F, F, F, F.*

   Test whether this arrangement of *A*'s and *F*'s may be regarded as random at 5% as well as at 10% level of significance.

8. Use a rank correlation at the 1% significance level and determine if there is significant positive correlation between the two samples on the basis of the following information:

Blender model	*A*1	*A*2	*A*3	*B*	*C*1	*C*2	*D*1	*D*2	*E*	*F*1	*F*2	*G*1	*G*2	*H*
Sample 1	1	11	12	2	13	10	3	4	14	5	6	9	7	8
Sample 2	4	12	11	2	13	10	1	3	14	8	6	5	9	7

9. A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure: 5, 2, 8, –1, 3, 0, –2, 1, 5, 0, 4, and 6. Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure?

An IQ test was administered to 5 persons before and after they were trained.

Candidates	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any improvement in IQ after the training program (test at 1% level of significance)

A state highway department has decided to investigate the severity of automobile accidents occurring at a particular urban intersection since the adoption of the right-turn-on-red law. From police records, they chose a random sample of ten accidents that occurred at the intersection before the law was enacted, and a random sample of ten accidents that occurred after the law was enacted. They used the total damage estimate for each accident as a measure of the accident's severity. The damage estimates are recorded in the table:

Before the law	150	500	250	301	242	435	100	402	716	200
After the law	145	390	680	560	899	1250	290	963	180	550

Test if the damages tended to increase after enactment of the law.

The production manager of a firm wants to examine the relationship between aptitude test scores given prior to hiring production-line workers and performance ratings received by the employees 3 months after starting work. The results of the study would allow the firm to decide how much weight to give to these aptitude tests relative to other work history information obtained, including references. The aptitude test results range from 0 to 100. The performance ratings are as follows: 1 = well below average, 2 = somewhat below average, 3 = average, 4 = somewhat above average, 5 = well above average. A random sample of 20 workers yielded the result listed here:

Employee	*Aptitude*	*Performance*
1	59	3
2	47	2
3	58	4
4	66	3
5	77	2
6	57	4
7	62	3
8	68	3
9	69	5
10	36	1
11	48	3
12	65	3
13	51	2
14	61	3
15	40	3
16	67	4
17	60	2
18	56	3
19	76	3
20	71	2

Can the manager infer at 5% level of significance that aptitude test scores are correlated with performance ratings?

14

Chapte

Linear Regression Analysis

In Chapter 8, we discussed about measuring the relationship between two characteristics or variat of a population based on paired samples collected on those variables. A good measure of relations between two variables is given by coefficient of correlation which tells us about the strength relationship and the direction of relationship as well. We also discussed that correlation coeffici measures only linear relationship between two variables. After determining the correlation betw two variables, we wish to determine a mathematical relationship between them so that we (i) predict the value of a variable based on the value the other variable, and (ii) explain the impac changes in the values of variable on the values of the other variable. Fitting a mathematical funct between two correlated variables using paired observations on them is studied in regression analy For mathematical convenience and ease, we fit a linear relationship. In this chapter, we discuss h to fit and analyze a regression model. We also discuss a generalized case of regression model wh we have many variables affecting the values of one variable.

14.1 DEPENDENT AND INDEPENDENT VARIABLES

Out of two variables, one is consider as a dependent variable and the other as independent variab Using regression model, we study those changes in the values of dependent variable that are result by the changes in the values of dependent variable. Or in other words, we regress the depende variable on the independent variable. Therefore, the dependent variable is also called as the regress variable or study variable, while the independent variable is called as regressor variable or explanato variable. Sometimes it is quite clear which one of the two variables is regressor and which one regressed. For instance, out of rainfall and yield of a certain crop, rainfall is independent variab which affects the yield of a certain crop. So, yield becomes dependent variable. In some situation it depends on the study. For instance, out of demand and supply, if you wish to study demand affected by supply, demand becomes regressed and supply becomes regressor. You have to take th other way in case you wish to study the supply as affected by demand.

2 SIMPLE LINEAR REGRESSION MODEL

ear regression model between a single study variable and a single explanatory variable is termed mple linear regression model. Later, in this chapter we discuss the models having a single study ble and many explanatory variables. Such models are termed as multiple linear regression models.

Let us denote the study variable by *Y* and explanatory variable by *X*. We collect paired observations $_1$), $(x_2, y_2), \ldots, (x_n, y_n)$ on the variables *X* and *Y*. These observations are plotted in the following ter diagram:

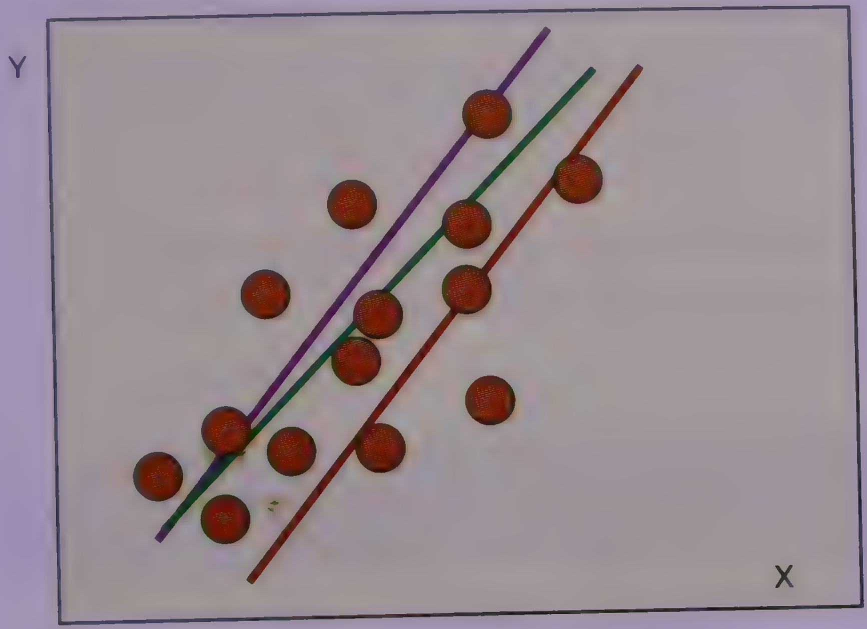


Fig. 14.1

In Fig. 14.1, the various dots together constitute the scatter diagram. The simplest relationship ween *X* and *Y* is a linear relationship which is given by *X* and $Y = a + bX$. Fitting a model means aining the values of the intercept term of the line (a) and slope of the line (b) on the basis of lected observations on *X* and *Y*. Different values of a and b will give different lines as shown in . 14.1. How to obtain the most suitable or the "best" values of a and b is a question which is cussed below:

.2.1 Least Squares Estimation

obtain some reasonably good estimate of a and b we use the method of least squares. We wish to ve such values of a and b which help us the most in studying *Y* as affected by *X*.

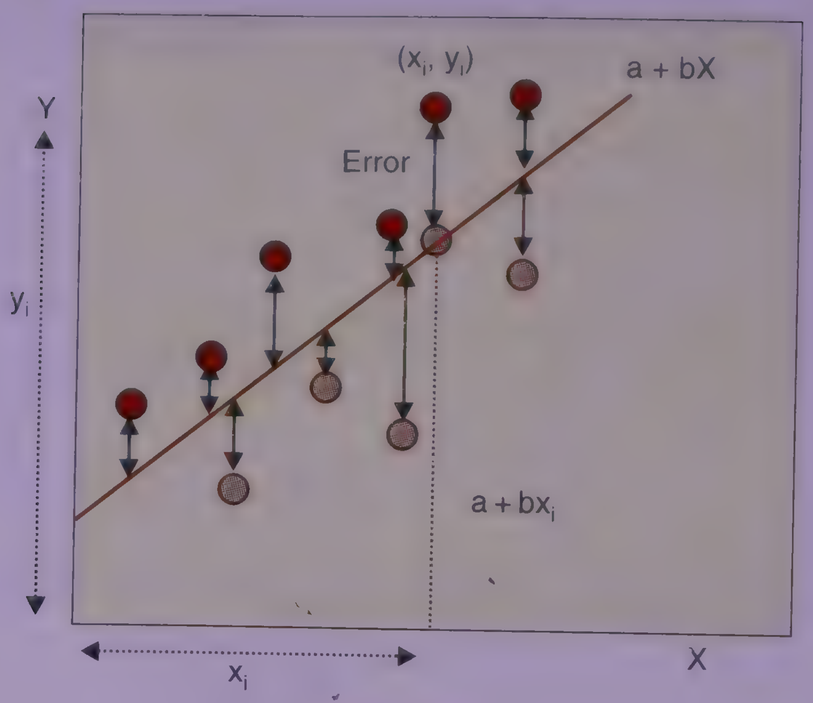


Fig. 14.2

In Fig. 14.2, we fit a line $Y = a + bX$. The exact relationship between X and Y is not linear, we only approximating the relationship by a line. Therefore, it is not correct to write the line equatio $Y = a + bX$. We write it as

$$\overline{Y} = a + bX$$

where $\overline{Y}$ is the predicted or fitted or estimated value of Y. The exact relationship between X an can be written as

$$Y = a + bX + \text{error}$$

The error in the above equation is shown in Fig. 14.2. This error is the difference between observed value and the predicted value of Y. Using collected observations (x_1, y_1), (x_2, y_2), (x_n, y_n), these errors or residuals can be written as $(y_i - a - bx_i)$ for $i = 1, 2, \ldots, n$. We wish to ha such values of a and b for which these residuals are minimum. In least squares method, we minim the summation of squared residuals. For this we differentiate $\sum_{i=1}^{n} (y_i - a - bx_i)^2$ with respect a and b separately and equate the derivatives to zero. Solving those two equations we get followi estimates of a and b:

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{SSXY}{SSX}$$

The values of a and b obtained using least squares method are called as least squares estimat (LSE) of a and b. Also, the relation between the correlation coefficient between X and

r) and LSE of b is given as

$$r = \hat{b}\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \hat{b}\sqrt{\frac{SSX}{SSY}}$$

Thus, the signs of r and $\hat{b}$ are the same.

Example 14.1: Watching television also reduces the amount of physical exercise, causing weight gains. A sample of fifteen 10-year old children was taken. The number of pounds each child was overweight was recorded (a negative number indicates the child is underweight). Additionally, the number of hours of television viewing per weeks was also recorded. These data are listed here.

TV	42	34	25	35	37	38	31	33	19	29	38	28	29	36	18
Overweight	18	6	0	–1	13	14	7	7	–9	8	8	5	3	14	–7

Fit the regression line and describe what the coefficients tell you about the relationship between the two variables.

Solution: We make the following table:

TV (x_i)	*Overweight* (y_i)	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
42	18	10.5333	12.2667	110.9511	129.2089
34	6	2.5333	0.2667	6.4178	0.6756
25	0	– 6.4667	– 5.7333	41.8178	37.0756
35	– 1	3.5333	– 6.7333	12.4844	– 23.7911
37	13	5.5333	7.2667	30.6178	40.2089
38	14	6.5333	8.2667	42.6844	54.0089
31	7	– 0.4667	1.2667	0.2178	– 0.5911
33	7	1.5333	1.2667	2.3511	1.9422
19	– 9	–12.4667	– 14.7333	155.4178	183.6756
29	8	– 2.4667	2.2667	6.0844	– 5.5911
38	8	6.5333	2.2667	42.6844	14.8089
28	5	– 3.4667	– 0.7333	12.0178	2.5422
29	3	– 2.4667	– 2.7333	6.0844	6.7422
36	14	4.5333	8.2667	20.5511	37.4756
18	–7	– 13.4667	– 12.7333	181.3511	171.4756
$\bar{x}$ = 31.4667	$\bar{y}$ = 5.7333	← Mean	Total →	SSX = 671.7333	$SSXY$ = 649.8667

Using the above calculation, we obtain $\hat{b}=\frac{SSXY}{SSX}=\frac{649.8667}{671.7333}=0.9674$ an

$\hat{a}=\bar{y}-\hat{b}\bar{x}=5.733-0.9674\times 31.4667=-24.709$. Therefore, the fitted simple linear regression mode is $\hat{y}=-24.709+0.9674\,x$

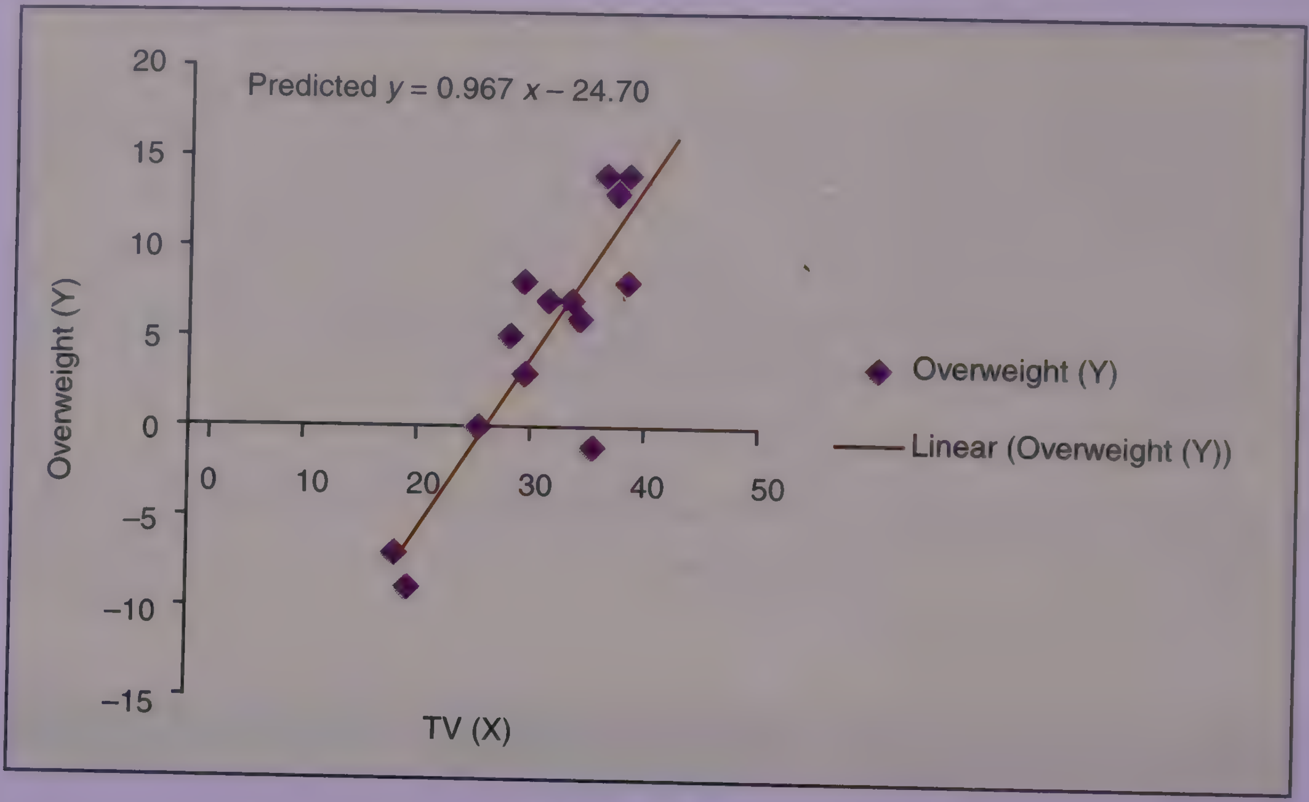


Fig. 14.3

Figure 14.3 shows the scatter plot of the data and the fitted line as well. The value $\hat{b}=0.9674$ is the change in the value of Y for a unit change in the value of X. The intercept is a constant or the value of Y when X is zero.

The fitted line can be used to predict some value of Y for a new (not in the collected data, however in the range) value of X. For instance, the predicted value of Y when $X = 30$, is $-24.709 + 0.967 \times 30 = 4.301$. Thus, weight gain for 30 hours of TV watching per week is 4.301 pounds.

It is advisable to predict Y values for those X values which are in the range of the collected sample as the behavior of the data may be different in future and past.

14.2.2 Coefficient of Determination

When the actual relationship between observed X and Y values is almost linear, the fitted linear model will be a reasonably good approximation of the actual relationship. In case, the actual relationship is not linear, the fitted linear model may be misleading. As discussed, the quantities $y_i - a - bx_i$ $(i = 1, 2, \ldots, n)$ are called as residuals. For a good model the magnitude of all the residuals should be as small as possible. Therefore, these residuals can be utilized to tell us something about the goodness of the fitted model.

Studying *Y* using *X* means explaining the variability of *Y* using *X*. Once we get the idea of how the values are changing, we can predict them. The variance of *Y* values is $\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$. We artition this variance into two parts as below:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

$$\text{or }\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

It can be shown that the product term $\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$. Since, $y_i-\hat{y}_i$ is the esidual, the quantity $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ is called as Sum of squares due to error (SSE). The quantity $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ is called as sum of squares due to regression (SSR) as it is the part of the variability of *Y* which is explained using regression model. Thus total variability in $Y\left(SST=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$ is partitioned into two parts — explained variability (SSR) and unexplained variability (SSE). Note that SST and SSY (discussed earlier) are the same. Clearly,

$$\text{SST} = \text{SSR} + \text{SSE}.$$

The fraction of SST explained by regression is given by R^2, i.e.,

$$R^2 = \text{SSR/SST} = 1 - (\text{SSE/SST})$$

It is clear that, $0 \le R^2 \le 1$. When SSR is closed to SST, R^2 will be closed to 1. This means that regression explains most of the variability in *Y* and the fitted model is good. When SSE is closed to SST, R^2 will be closed to 0. This means that regression does not explain much variability in *Y* and the fitted model is not good. The quantity R^2 is called as coefficient of determination and is used to evaluate the goodness of the fitted model. For simple linear regression model, coefficient of determination R^2 is the same as square of correlation coefficient between *X* and *Y*.

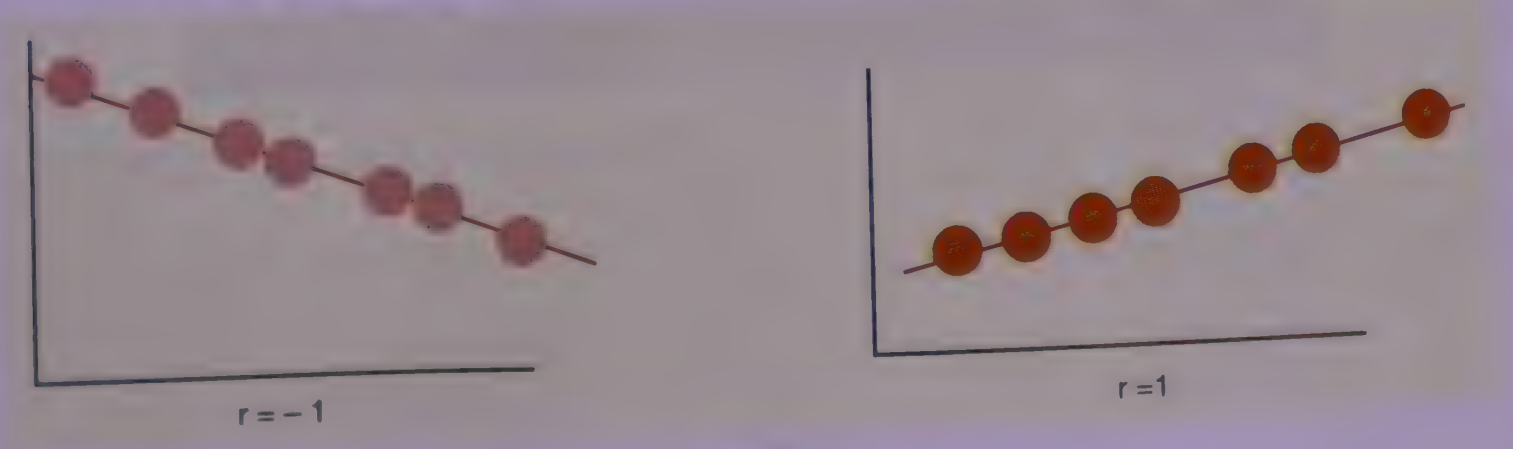


Fig. 14.4(a)

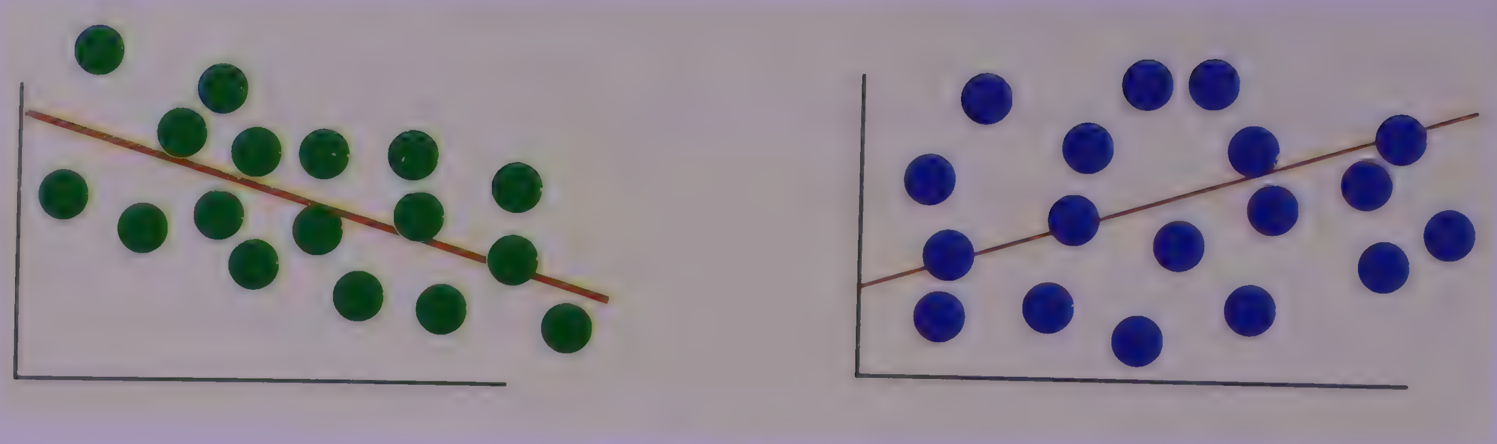


Fig. 14.4(b)

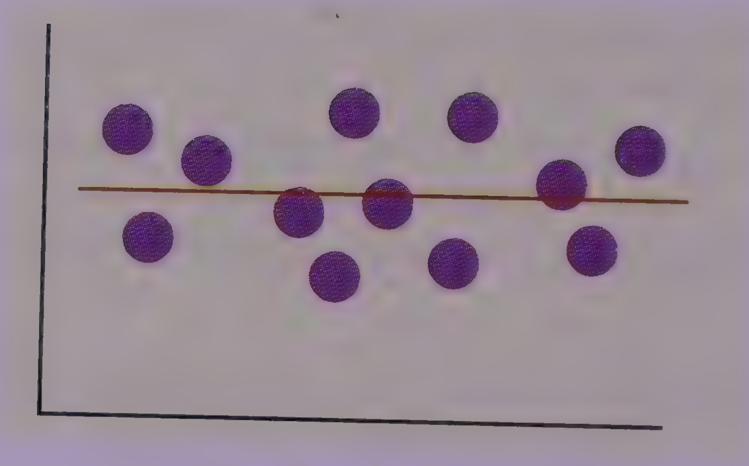


Fig. 14.4(c)

In Fig. 14.4(a), $R^2 = 1$, which means a perfect linear relationship between X and Y. In these models, 100% of the variation in Y is explained by X. In Fig. 14.4(b), $0 < R^2 < 1$, which indicates relatively weak linear relationships. In this case, some but not all of the variation in Y is explained by X. In Fig.14.4(c), $R^2 = 0$, which means no linear relationship. In this model, none of the variation in Y is explained by X.

Example 14.2: Obtain R^2 for the model fitted in Example 14.1 and comment on its value.

Solution: Using the data and the fitted model in Example 14.1, we make the following table:

y_i	$\hat{y}_i = \hat{a} + \hat{b}x_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
18	15.9218	4.318915	150.4711
6	8.1826	4.763743	0.071111
0	– 0.524	0.274576	32.87111
– 1	9.15	103.0225	45.33778
13	11.0848	3.667991	52.80444

y_i	$\hat{y}_i = a + bx_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
14	12.0522	3.793925	68.33778
7	5.2804	2.957024	1.604444
7	7.2152	0.046311	1.604444
– 9	– 6.3284	7.137447	217.0711
8	3.3456	21.66344	5.137778
8	12.0522	16.42032	5.137778
5	2.3782	6.873835	0.537778
3	3.3456	0.119439	7.471111
14	10.1174	15.07458	68.33778
– 7	– 7.2958	0.087498	162.1378
		SSE = 190.2216	SST = 818.9333

Thus, $R^2 = SSR/SST = 1 - SSE/SST = 1 - 190.2216/818.9333 = 0.7677$ which is close to 1. Thus, he fitted model is considered to be good one.

4.2.3 Standard Error

ll the observations of a data-set cannot be exactly the same as their arithmetic mean (AM). Variability f the observations around AM is measured by standard deviation. Similarly in regression, all Y alues cannot be the same as predicted Y values. Variability of Y values around the prediction line is neasured by standard error of the estimate which is given by $S_{YX} = \sqrt{\dfrac{SSE}{n-2}}$. When the predicted alues and observed values are close, standard error is small. However, standard error is not a very ood measure of judging the goodness of the fitted model. It should be considered along with coefficient f determination and other measures to be discussed later.

14.2.4 Assumptions or Conditions Required

The relationship between X and Y is almost linear. There is no meaning of approximating a non-linear elationship by a line.

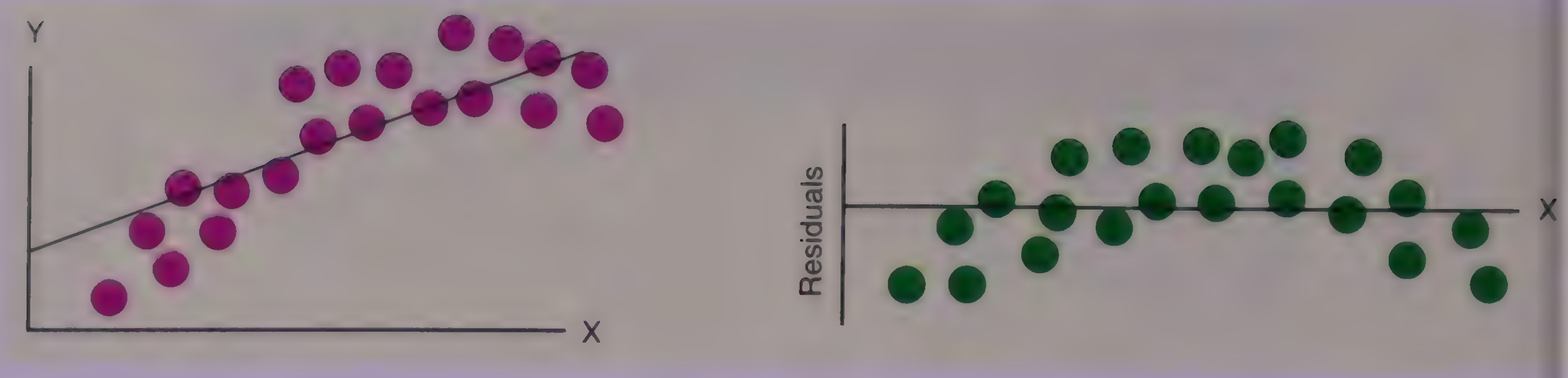


Fig. 14.5(a)

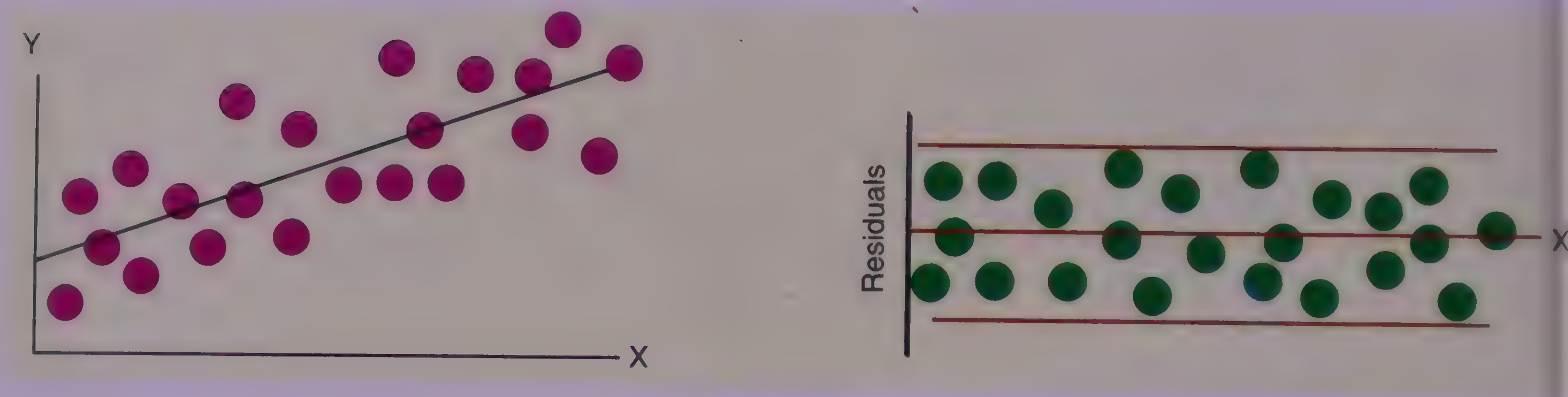


Fig. 14.5(b)

Figure 14.5(a) shows non-linear relationship, while Fig. 14.5(b) shows linear relationship.

Errors or residuals are statistically independent. This can be made sure by collecting a rando sample, i.e., independent pairs of observations in the sample.

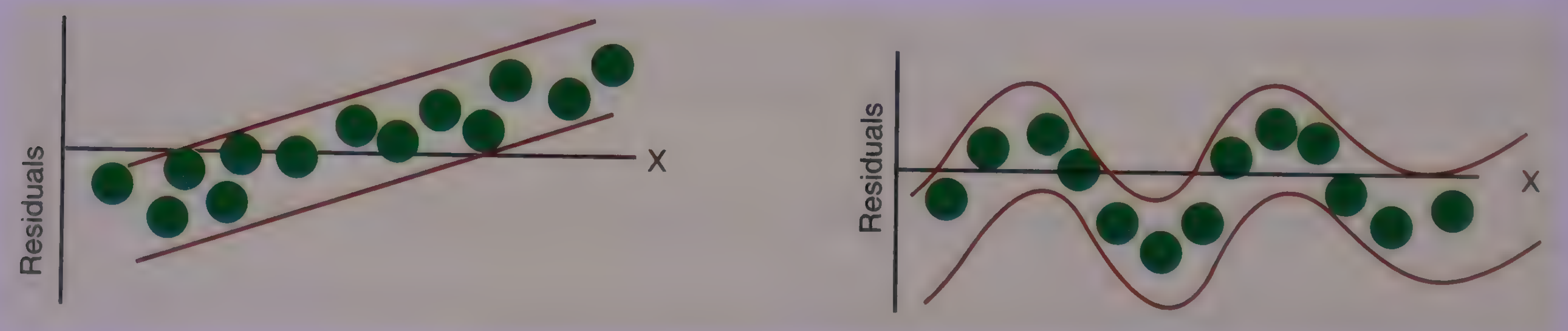


Fig. 14.6(a)

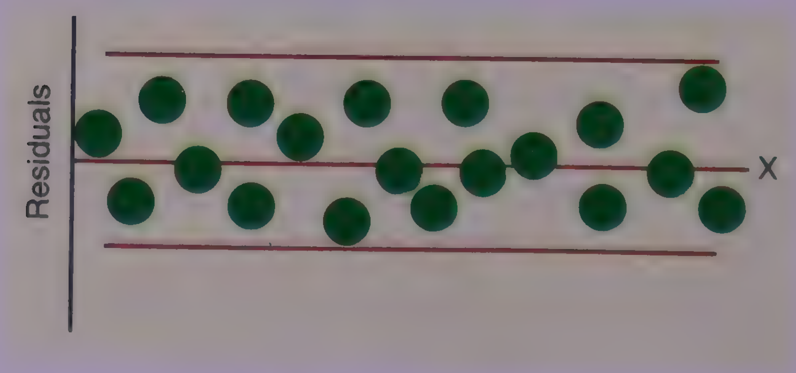


Fig. 14.6(b)

In Fig. 14.6(a), we can see the situation when the observations or the errors are not independent. hile in Fig. 14.6(b) we see the scatter plot between residuals and X values when the errors are dependent or uncorrelated to be more precise. The assumption of independence can also be examined ing Durbin Watson statistic which is given by

$$D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

here $e_i = y_i - a - bx_i$ $(i = 1, 2, \ldots, n)$ are residuals. The value of D lies between 0 and 4. When it is ear 2, residuals are uncorrelated. For positively correlated residuals, D approaches 0 and for negatively orrelated residuals, it approaches 4. In practice a value of D between 1 and 3 is usually considered be accepted.

All the errors have a common variance (Homoscedasticity). When all the error variances are ot the same, we have heteroscedastic models. Such models are fitted using some advanced method.

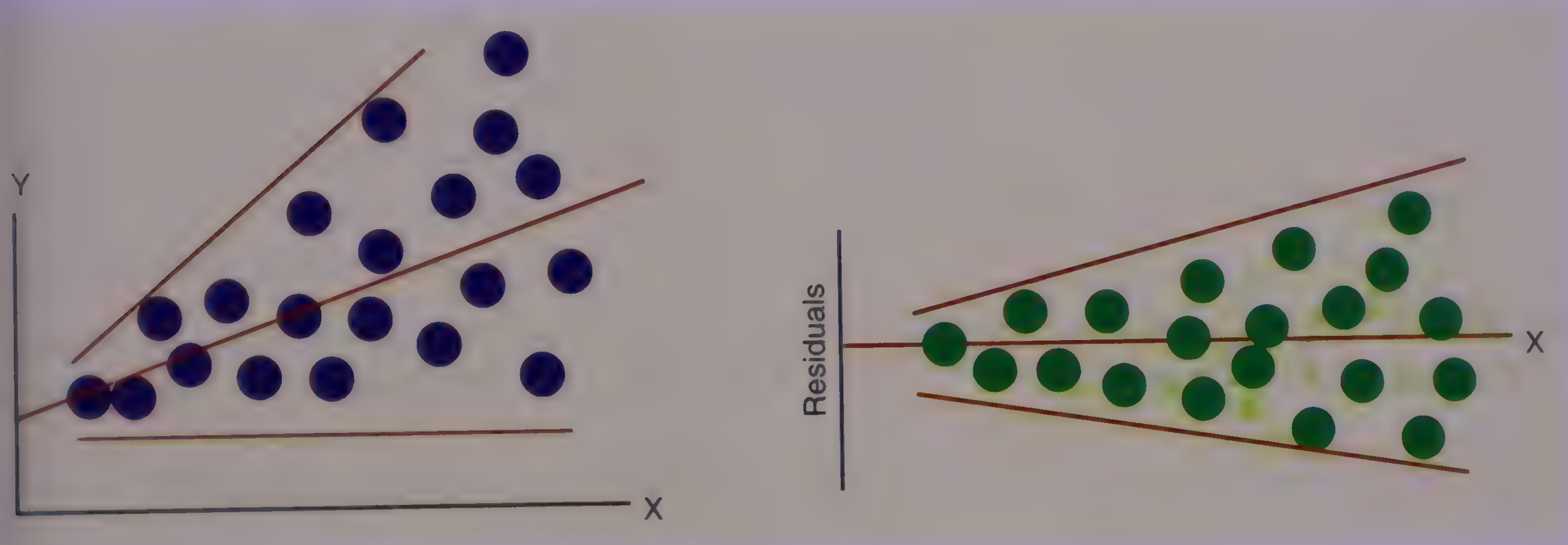


Fig. 14.7(a)

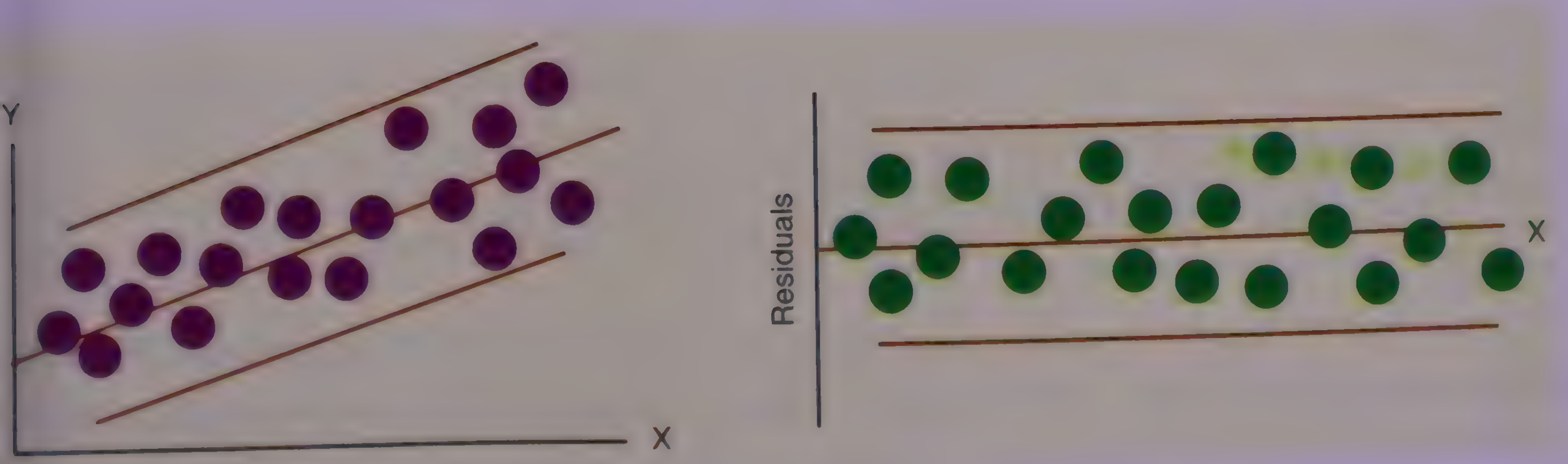


Fig. 14.7(b)

In Fig. 14.7(a), we see a heteroscedastic model, while Fig. 14.7(b) shows a homoscedas model.

Example 14.3: Obtain the standard error of estimate and examine the assumptions of lineari independent errors, and homoscedasticity for the model fitted in Example 14.1.

Solution: Standard Error of estimate in Example 14.1 is $S_{YX} = \sqrt{\frac{SSE}{n-1}} = \sqrt{\frac{190.2216}{15-2}} = 3.8252$

Standard error can take any positive value. We can compare standard errors of two mode However, making any conclusion using a single model's standard error is not meaningful.

Assumption 1: Linearity

The scatter plot of the data is presented in Figure 14.3 which suggests almost linear relationship. W below plot the scatter diagram between X values and residuals.

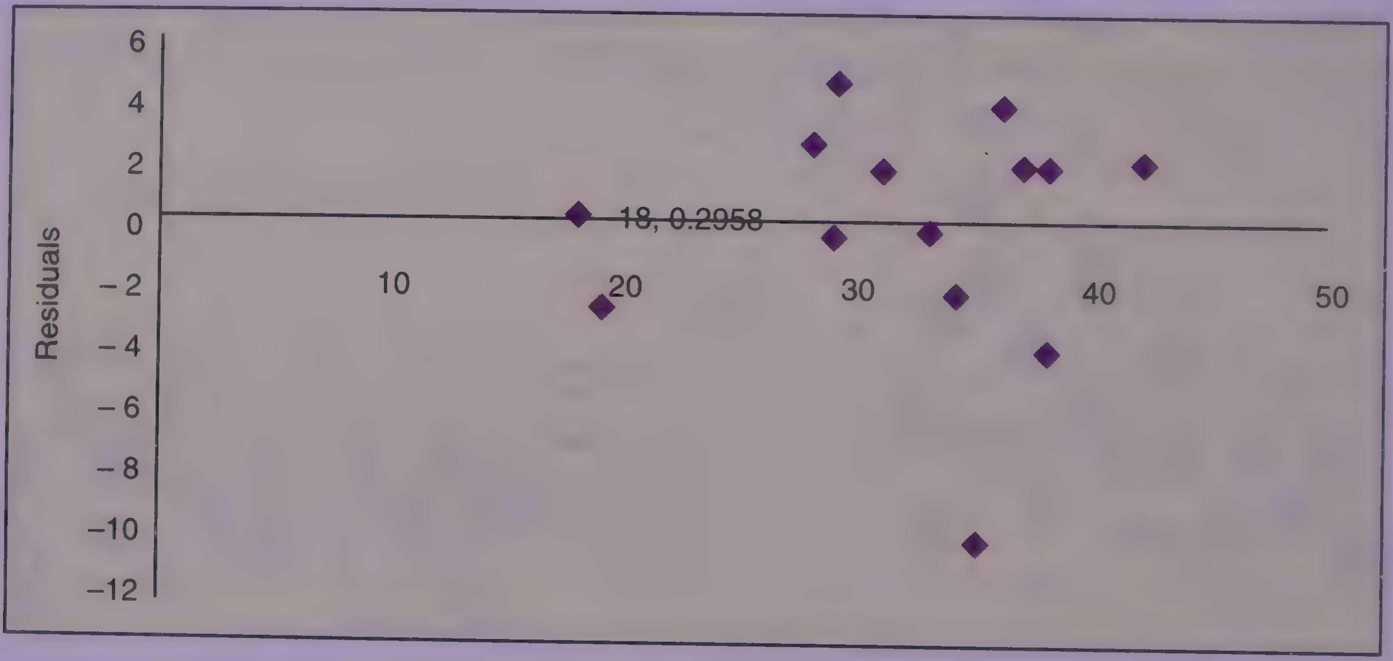


Fig. 14.8

Other than one outlier, all other data-points are satisfying the linearity assumption.

Assumption 2: Independence

From Fig. 14.8, it is clear that the given data satisfies the independent errors assumption also.

The Durbin Watson statistic is $D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = 2.6726$.

The value of D close to 2 implies uncorrelated residuals. The values between 1 and 3 are als accepted, while the values 0 to 1 and 3 to 4 are alarming. Thus the residuals are almost uncorrelate

Assumption 3: Homoscedasticity

From Fig. 14.3 and 14.8, homoscedasticity of the model is also verified.

.2.5 Testing the Hypothesis about the Slope and Correlation Coefficient

this section we conduct hypothesis testing about the linear relationship between the variables der study.

.2.5.1 t-Test for the Slope

the model $Y = a + b\,X + \text{error}$, if $b = 0$, then the model cannot be considered as a linear model. herefore, here we test H_0: $b = 0$ against $H_a : b \neq 0$. The test statistic is

$$T_c = \frac{\hat{b}}{\sqrt{\dfrac{SSE}{(n-2)SSX}}}$$

here the notations have their usual meanings. When the distributions of errors is normal, the test atistic T_c follows Student's *t* distribution with $(n-2)$ degree of freedom. Note that, until here we do ot require normality assumption. Here the normality is required for the validity of *t* distribution. For large sample normality can be examined by plotting the histogram. For small samples, it is difficult verify the normal distribution of residuals. However, Q-Q plot may be helpful up to some extent.

As clear from the alternative hypothesis, it is a two-tailed test. Therefore we reject null hypothesis t $\alpha \times 100\%$ level of significance, when the computed value T_c is lower than $-t_{\alpha/2}$ or larger than $t_{\alpha/2}$ or when $|T_c| > t_{\alpha/2}$). Here $t_{\alpha/2}$ is the critical value obtained from *t* table. Rejecting the null hypothesis eans that there is an evidence of linear relationship.

xample 14.4: For the model fitted in Example 14.1, test the hypothesis that the slope is zero gainst it is not zero (examine the linearity).

olution: We wish to test here H_0: $b = 0$ against $H_a : b \neq 0$

The test statistic is

$$T_c = \frac{\hat{b}}{\sqrt{\dfrac{SSE}{(n-2)SSX}}} = \frac{0.9674}{\sqrt{\dfrac{190.2216}{(15-2)\times 671.7333}}} = 6.5546$$

At 5% level of significance and 13 d.f., the critical value using *t* distribution is 2.160 which is maller than the computed value. Therefore, at 5% level of significance we reject the null hypothesis nd conclude that there is an evidence of linear relationship.

14.2.5.2 F-Test (ANOVA)

-test can also be used as an alternative to *t*-test for linearity. Here also we test H_0: $b = 0$ against $H_a : b \neq 0$. This test is based on the concept of partitioning the total variability *SST* into explained ariability *SSR* and unexplained variability SSE. We have already discussed the formulae of obtaining ST, *SSR* and *SSE*. After obtaining these sums of squares we make following ANOVA table:

ANOVA	d.f.	Sum of Squares	Mean Sum of Squares	F-ratio
Regression or Explained	1	SSR	$MSR = SSR$	$F_c = MSR/MSE$
Residual or Unexplained	$n-2$	SSE	$MSE = SSE/(n-2)$	
Total	$n-1$	SST		

Under the normality of residuals, the test statistic F_c follows Snedecor's F distribution wi $(1, n-2)$ d.f. We reject the null hypothesis for this right tailed test (ANOVA is a right tailed metho when the computed value is higher than the critical value F_α.

Example 14.5: For the model fitted in Example 14.1, test the hypothesis that the slope is ze against it is not zero (examine the linearity) using ANOVA method.

Solution: We wish to test here H_0: $b = 0$ against $H_a : b \neq 0$.

ANOVA	d.f.	Sum of Squares	Mean Sum of Squares	F-ratio
Regression or Explained	1	628.7118	628.7118	$F_c = 42.96702$
Residual or Unexplained	13	190.2216	14.63243	
Total	14	818.9333		

The computed test statistic $F_c = 42.96702$. At 5% level of significance and (1, 13) d.f., the critic value using F distribution is 4.67 which is smaller than the computed value. Therefore, at 5% level significance we reject the null hypothesis and conclude that there is an evidence of linear relationshi

Note that, because of the relationship of t and F distributions, the critical value of F, 4.67, is th square of the critical value of t, 2.160. Same relationship can be observed in the computed values

14.2.5.3 t-Test for correlation coefficient

When the variables X and Y are (linearly) correlated, it is meaningless to fit a linear regression mode between them. Therefore, we may like to examine whether there is some significant linear relationshi between X and Y or not. Here we wish to test H_0: 'correlation coefficient between X and Y is zer against H_a: 'correlation coefficient between X and Y is not zero'. The test statistic is

$$T_r = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

where, r is the correlation coefficient between the sample observations of the variables X and Y. Th formula of obtaining correlation coefficient has already been discussed in Chapter 8. A relationshi etween correlation coefficient and *LSE* of slope is given in Section 14.2.1. When the residuals have ormal distribution, the distribution of this test statistic T_r is Student's *t* with (n–2) d.f. We take our ecision about the null hypothesis in this two-tailed by comparing the computed value and critical alue of *t* distribution. We reject null hypothesis at α × 100% level of significance, when the computed alue T_r is lower than $-t_{\alpha/2}$ or larger than $t_{\alpha/2}$ (or when $|T_r| > t_{\alpha/2}$). Here $t_{\alpha/2}$ is the critical value btained from *t* table. Rejecting the null hypothesis means that there is a significant linear relationship etween the variables.

xample 14.6: For the model fitted in Example 14.1, test the hypothesis that the correlation coefficient etween TV (X) and Overweight (Y) is zero or not.

olution: We wish to test H_0: 'correlation coefficient between X and Y is zero' against H_a: 'correlation oefficient between X and Y is not zero'.

Correlation coefficient in sample values is $r = \hat{b}\sqrt{\dfrac{SSX}{SSY}} = 0.8762$

The test statistic is $T_r = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \dfrac{0.8762}{\sqrt{\dfrac{1-0.8762\times 0.8762}{15-2}}} = 6.5550$

At 5% level of significance and 13 d.f., the critical value using *t* distribution is 2.160 which is smaller than the computed value. Therefore, at 5% level of significance we reject the null hypothesis and conclude that there is significant correlation between the variables.

14.2.6 Confidence Interval Estimate of the Slope

The point estimation of the slope b is given by the least squares estimate $\hat{b} = SSXY / SSX$ A confidence interval estimate of b at a given confidence coefficient (1–α) can be obtained by

$$\hat{b} \mp t_{\alpha/2}(d.f.=n-2)\times\sqrt{\frac{SSE}{(n-2)SSX}} \quad \text{or} \quad \hat{b} \mp t_{\alpha/2}(d.f.=n-2)\times\sqrt{\frac{MSE}{SSX}}$$

Confidence interval estimate of slope b is not much used in practice.

Example 14.7: Construct a 95% confidence interval estimate of slope b for the model fitted in Example 14.1.

Solution: For 95% confidence coefficient and 13 d.f., the critical value using *t* distribution is 2.160. Therefore, 95% confidence interval estimate of slope b is given by

$$\hat{b} \mp t_{\alpha/2}(d.f.=n-2)\times\sqrt{\frac{MSE}{SSX}} \quad \text{or} \quad 0.9674 \mp 2.160\times\sqrt{\frac{14.63242}{671.7333}} \quad \text{or } (0.6486,\ 1.2862).$$

14.2.7 Confidence Interval

In Section 14.2.1, we used the fitted line to predict the Y value for $X = 30$. Weight gain for 30 hou of TV watching per week is estimated as 4.301 pounds. This is a point estimate of the populati mean (of Y population) at a given X value. We can also obtain the confidence interval estimate population mean or mean response at some given X value as

$$\hat{y}_i \mp t_{\alpha/2}(d.f.=n-2)S_{XY}\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{SSX}}$$

where, $S_{YX}=\sqrt{\frac{SSE}{n-2}}$ is the standard error of the estimate.

The width of the confidence interval depends on the confidence coefficient α, standard error o the estimate, and the given X value at which we are obtaining the confidence interval. When the values are close to the mean $\bar{x}$, the width of the interval is smaller. The width increases when th deviation of X values increases from their mean.

For Example 14.1, the confidence interval and the predicted Y values at each X value in th sample are shown in the following Fig. 14.9.

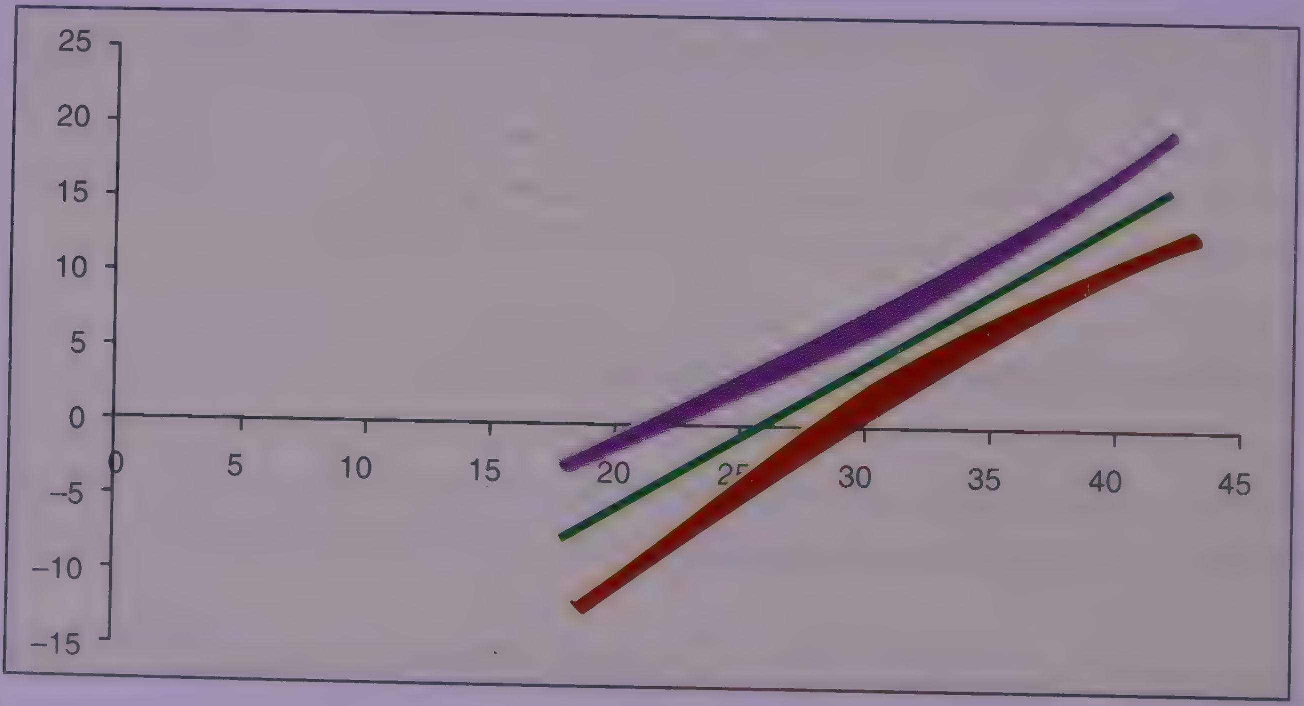


Fig. 14.9

The upper line represents the upper confidence limits; lower line shows the lower confidence limits; and the middle line represents the predicted Y values.

14.2.8 Prediction Interval

Using confidence interval in Section 14.2.7, we estimated the mean value of Y or mean response which is a parameter. In order to predict the individual value of study variable (or the response) fo a given X value, we obtain prediction interval as

$$\hat{y}_i \mp t_{\alpha/2}(d.f.=n-2)S_{YX}\sqrt{1+\frac{1}{n}+\frac{(x_i-\overline{x})^2}{SSX}}$$

xample 14.8: Using the model fitted in Example 14.1, construct a 95% prediction interval for edicting the overweight for a child watching television for 30 hours per week.

olution: 95% prediction interval of overweight for $X = 30$ is

$$\hat{y}_i \mp t_{\alpha/2}(d.f.=n-2)S_{YX}\sqrt{1+\frac{1}{n}+\frac{(x_i-\overline{x})^2}{SSX}} \text{ or } 4.301 \mp 2.160\times 3.8252\sqrt{1+\frac{1}{15}+\frac{2.1511}{671.7333}} \text{ or}$$

-4.2452, 12.8472).

4.2.9 Using Excel

e can use 'Data Analysis' tool pack (Add-in) in MS Excel. If you have MS Excel 2003 or advance ersion, you can download and install it for free. In 'Excel options', select 'Add-Ins' and then 'Excel dd-Ins'. In the pop-up window, check 'Analysis ToolPack' and 'Analysis ToolPack – VBA' and lick OK. Make sure that your computer is connected to the internet.

Summary Output

Regression Statistics	
Multiple R	0.876196547
R Square	0.767720389
Adjusted R Square	0.749852726
Standard Error	3.825235201
Observations	15

- 'Multiple R' is the correlation coefficient or the square root of R^2.
- 'R square' is R^2.
- 'Adjusted R Square' is of no use in simple linear regression models. It will be discussed later in multiple models.
- 'Standard Error' is the standard error of the estimate.
- 'Observations' gives number of data points.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	628.7118	628.7118	42.96703	1.83945E-05
Residual	13	190.2215	14.63242		
Total	14	818.9333			

This is the p-value of the test. Since it is less than 0.05, so the null hypo-thesis is rejected and we conclude that there is an evidence of linear relationship

LSE of intercept and slope

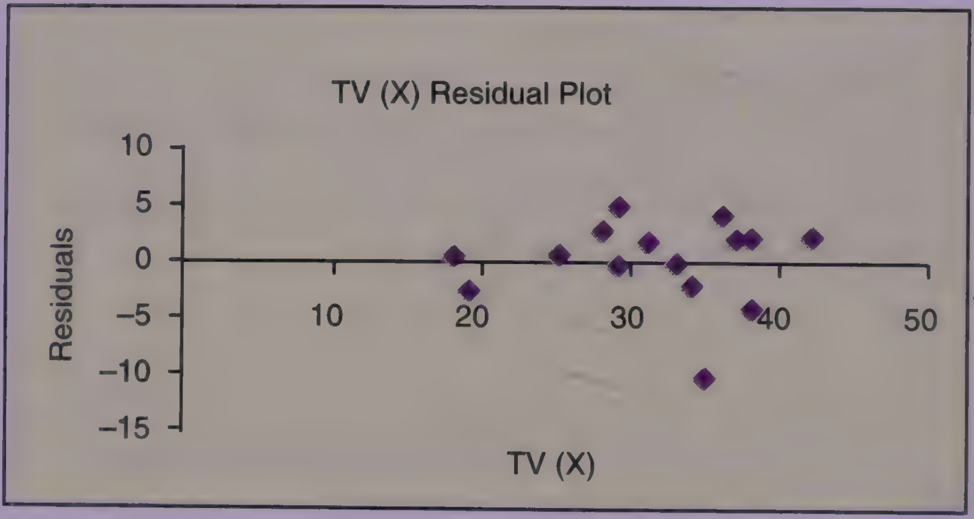
Standard errors of LSE.

Computed *t* statistic for intercept and slope.

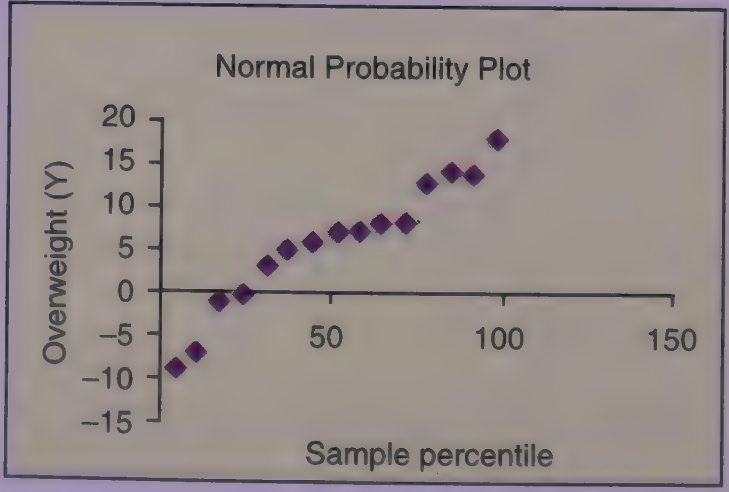
p-value for intercept and slope. Both are below 0.05. So, slope is not zero and intercept is not zero as well at 5% level.

	Coefficients	*Standard Error*	*t-State*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	–24.7090	4.7481	–5.2040	0.0002	–34.9666	–14.4515
TV (X)	0.9674	0.1476	6.5549	0.0000	0.6486	1.2863

Confidence interval estimate of intercept and slope.



Scatter plot between x values and residuals used to verify the assumptions of linearity, homoscedasticity, and independence.



Normal Probability Plot or Q-Q plot is constructed to verify normality of the data. In normal sample, we get a plot almost making a line at 45 degree from x-axis. This plot shows almost normal distribution.

14.3 MULTIPLE LINEAR REGRESSION MODEL

In simple linear regression analysis, we fit linear relation between one explanatory variable (X) an one response variable (Y). In some situations, the response variable Y may depend on more than on xplanatory variables. For example, cost of certain item may depend on (i) labor cost, (ii) electricity ost, and (iii) raw material cost. Salary of a person may depend on his or her (i) education and i) experience. Sales of certain item may depend on (i) cost and (ii) advertising expenditure.

Suppose that the response variable (Y) depends on k explanatory variables denoted as $_1, X_2, \ldots, X_k$. The linear relationship of Y as a function of $X_1, X_2, \ldots, X_k$ can be written as

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \text{error}.$$

Above model is called as multiple linear regression model. In this model, a is the intercept or onstant term and $b_1, b_2, \ldots, b_k$ are called as regression slopes or regression coefficients. In order to it this model, we need to estimate the regression coefficients $a, b_1, b_2, \ldots, b_k$. For this, we use the nethod of least squares.

14.3.1 Least Squares Estimation

As in simple linear regression, we minimize the sum of squares due to residuals, $S(a, b_1, b_2, \ldots, b_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ with respect to $a, b_1, b_2, \ldots, b_k$. For this, we differentiate S with respect $a, b_1, b_2, \ldots, b_k$ to separately and equate the derivative to zero. On solving those normal equations, we get the least squares estimates of regression coefficients. Let us represent the regression coefficients in vector form as

$$b = \begin{pmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix}$$

The least squares estimate of b is given by

$$\hat{b} = \begin{pmatrix} \hat{a} \\ \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_k \end{pmatrix} = (X'X)^{-1}X'y$$

where X is the matrix of order $n \times (k+1)$, each row representing a data-point, each column (from 2nd to $(k+1)$th) representing the observations on explanatory variables $X_1, X_2, \ldots, X_k$, first column having 1's as its elements. y is a vector of the n values of response variable Y. The elements of vector $\hat{b}$ gives LSE of regression coefficients $a, b_1, b_2, \ldots, b_k$. Thus, predicted response:

$$\hat{Y} = \hat{a} + \hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_kX_k$$

It may be time taking to solve the matrix formula of LSE. Therefore, we use MS Excel to obta LSE of regression coefficients in the following example.

Example 14.9: A distributor of frozen dessert pies wants to evaluate factors which influence th demand. He considers Pie sales (units per week) as the response variable, Price (in USD) as o explanatory variable, and Advertising Expenditure (in 100 USD) as another explanatory variabl Data on these three variables are collected for 15 weeks:

Week	*Pie Sales*	*Price (in USD)*	*Adv. Ex (in 100 USD)*
1	350	5.5	3.3
2	460	7.5	3.3
3	350	8	3
4	430	8	4.5
5	350	6.8	3
6	380	7.5	4
7	430	4.5	3
8	470	6.4	3.7
9	450	7	3.5
10	490	5	4
11	340	7.2	3.5
12	300	7.9	3.2
13	440	5.9	4
14	450	5	3.5
15	300	7	2.7

Fit a regression model to predict Pie Sales using Price and Advertising Expenditure.

Solution: As discussed in simple linear regression, we use 'Data Analysis' tool pack of MS Excel t fit the regression model. Below, we present the last table of the Excel outcome partially:

	Coefficients
Intercept	306.5261933
Price	–24.97508952
Adv. Ex.	74.13095749

In this table, we get the LSE of regression coefficients. Thus the fitted model is

Predicted Pie Sales = 306.53 – 24.98 Price + 74.13 AdvExp, where 'Price' means the explanator variable 'Price', and 'AdvExp' means the explanatory variable 'Advertising Expenditure'. Th interpretation of regression coefficients is as follows. Pie sales will decrease, on average, by 24.98

roughly 25) pies per week for each 1 USD increase in selling price, while advertising expenses e kept fixed. Pie sales will increase, on average, by 74.13 (or roughly 74) pies per week for each 0 USD increase in advertising, while selling price are kept fixed.

Fitted model can be used for prediction purpose. Suppose, we wish to predict oie sales in the eek when selling price is 5.50 USD and advertising expenditure is 350 USD. Predicted pie sales in at week = 306.53 – 24.98 (5.50) + 74.13 (3.5) = 428.62 pies or roughly 429 pies.

4.3.2 R² and Adjusted R² Coefficients

s in simple linear regression analysis, here also we can use coefficient of determination R^2 to judge oodness of fitted model by the same formula:

$$R^2 = SSR/SST = 1 - (SSE/SST),$$

here SSR, SSE, and SST are obtained using the same formula as discussed earlier. The only fference is in calculating predicting Y values which are given by $\widehat{Y} = \hat{a} + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \ldots + \hat{b}_k X_k$.

In Example 14.8, $R^2 = 0.5215$. If we consider fitting a simple linear regression model taking Pie ales as response variable and Price as explanatory variable, $R^2 = 0.1965$. If we fit a simple linear gression model taking Pie Sales as response variable and Advertising Expenditure as explanatory riable, $R^2 = 0.3095$.

Thus, we observe here R^2 is higher when there are more explanatory variables in the model. This not just a coincidence, but a property of R^2. Whenever an explanatory variable is added to the odel, the value of R^2 increases. This increase is regardless of the contribution of newly added xplanatory variable. Thus the value of R^2 may be misleading in multiple linear regression models. herefore, an adjusted value of R^2 is defined, which is called as adjusted R^2 and defined as

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Adjusted R^2 increases only if the added explanatory variable contributes in explaining the variation response variable Y.

In Example 14.8, Adjusted $R^2 = 0.4417$.

4.3.3 Standard Error

tandard error of the estimate for a multiple linear regression model is given by

$$S_{YX} = \sqrt{\frac{SSE}{n-k-1}}$$

Standard error of the estimate in Example 14.8 is 47.4634.

14.3.4 ASSUMPTIONS

AS in simple linear regression analysis, in multiple linear regression analysis, following assumpti are required:

(i) Linearity

(ii) Independence

(iii) Homoscedasticity

We have more than one (k) explanatory variables, so we need to make k scatter plots Y with Y with X_2, ..., Y with X_k. In case, individual relationship of each explanatory variable with respo variable is linear, the overall relationship of response variable with all explanatory variables ta together may not be linear. Therefore, individual scatter plots may not always help in getting the i of overall relationship. In order to verify linear relationship, we plot a scatter plot of residuals w predicted Y values. In Fig. 14.10, the left side figure represents a nonlinear relationship while the ri side figure represents a linear relationship.

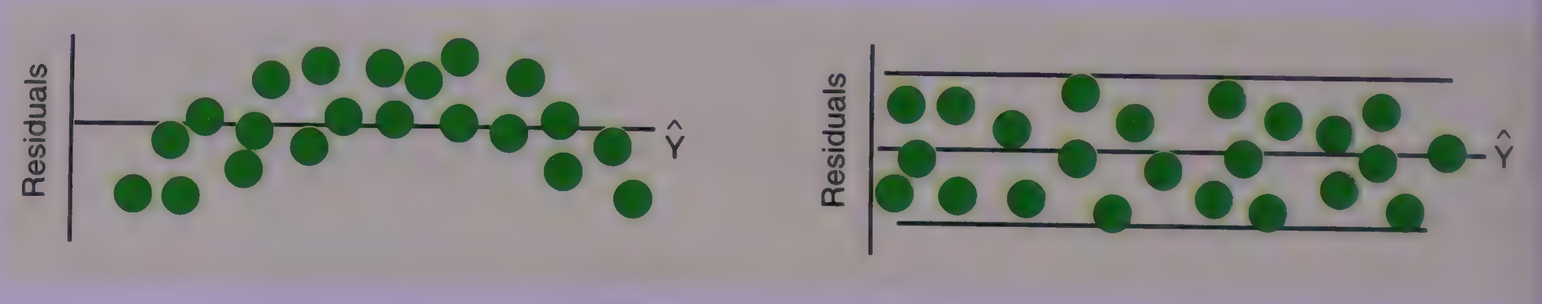


Fig.14.10

In order to examine independence of errors, we again use the same scatter plot. In Fig. 14. the first and second figures from left side represent correlated residuals while the extreme right si figure shows uncorrelated residuals.

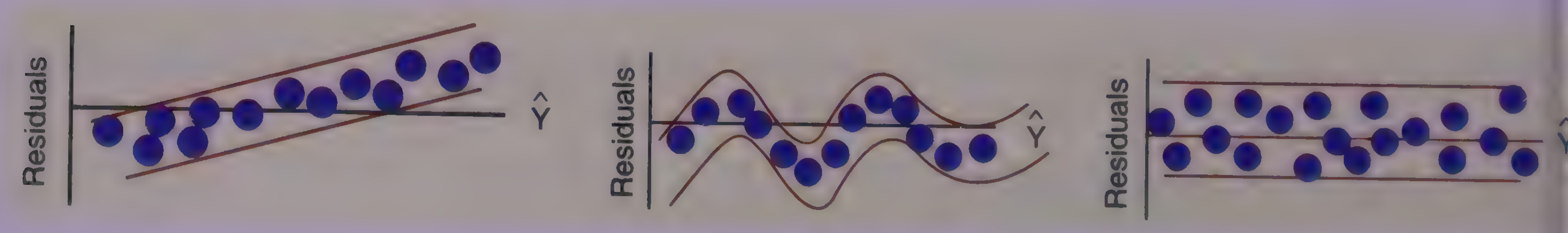


Fig. 14.11

Durbin Watson statistic can also be obtained to examine the assumption of independence usi the formula discussed in Section 14.2.4.

The assumption of homoscedasticity (all error variances are the same) can also be examined using the scatter plot of residuals and predicted Y values. In Figure 14.12, left side figure shows heteroscedastic model, while right side figure shows a homoscedastic model.

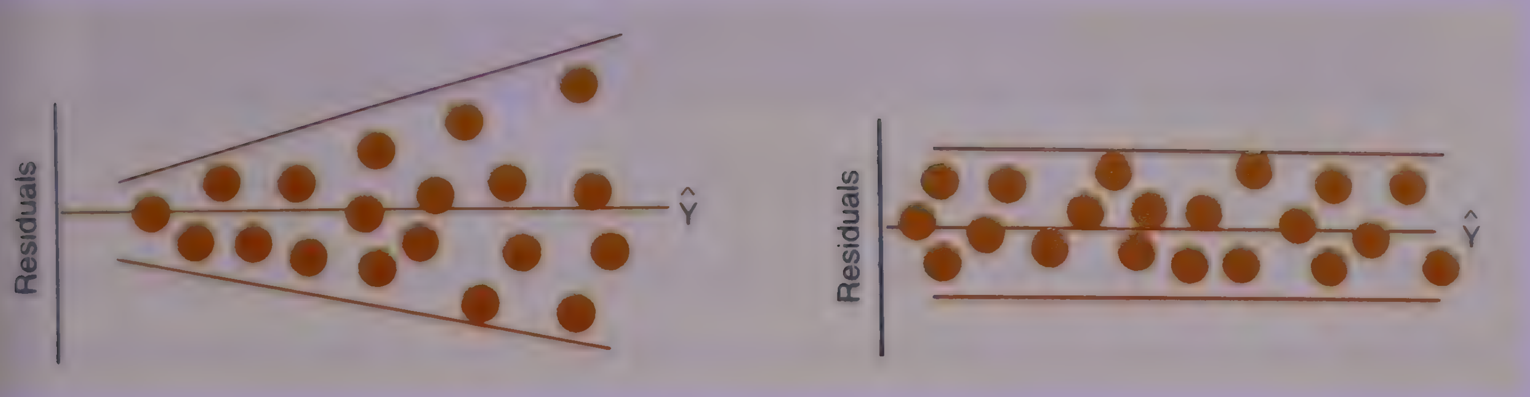


Fig. 14.12

Along with these three conditions, we also require the number of observations (n) to be greater an the number of explanatory variables (k), i.e., $n > k$. When n is not larger than k, we are not be le to obtain the estimates of regression coefficients and so we cannot fit the regression model.

We also assume that the explanatory variables are statistical independent of one another. In case s assumption is not satisfied, we face the problem of multicollinearity discussed later.

.3.5 *t*-Test for the Slopes

is is also called as test of hypothesis for individual explanatory variable. We test for each explanatory riable X_j ($j = 1, 2, \ldots, k$), the following hypothesis: H_0: $b_j = 0$ (no linear relationship between X_j d Y) against $H_a : b_j \neq 0$ (linear relationship exists between X_j and Y).

As in simple linear regression model, we use a two tailed t-test. If the null hypothesis is accepted, means that there is no evidence of linear relationship between X_j and Y and the variable X_j may be leted from the model.

The test statistic:

$$T_c = \frac{\hat{b}_j}{\text{Standard Error}(\hat{b}_j)} = \frac{\hat{b}_j}{\sqrt{MSE \times C_{jj}}}$$

here C_{jj} is the (j, j)th element of matrix $(X'X)^{-1}$ and MSE = SSE/(n–k–1) is called as mean uared error or mean sum of squares due to residuals. MSE provides a feasible estimate of the mmon error variance. The test statistic T_c follows Student's t with (n–k–1) degree of freedom ovided the distribution of residuals is normal.

Because of the complicated matrix calculations, we solve the following example using MS Excel.

xample 14.10: In Example 14.8, test the hypothesis about individual explanatory variable and ggest if any one of the two explanatory variables should be removed from the model.

lution: As discussed in simple linear regression, we use 'Data Analysis' tool pack of MS Excel to the regression model. Below, we present the last table of the Excel outcome partially:

	Coefficients	*Standard Error*	*t-Stat*	*P-value*
Intercept	306.5261933	114.2539	2.682851	0.019932
Price	–24.97508952	10.83213	–2.30565	0.039788
Adv. Ex.	74.13095749	25.96732	2.854779	0.014494

In first column of this table we are given the explanatory variables (Price and AdvEx). In t second column, we ghet the least squares estimate of Intercept, Price, and AdvEx, respective Third column gives the standard error of each least square estimate. In fourth column, we get t computed test statistic (T_c) obtained by dividing the least square estimates (*Coefficients*) corresponding standard error. Last column provides the *p*-value the test of each individual explanatc variable as well as the intercept. As we know, we compare *p*-value with some level of significan and reject the null hypothesis if *p*-value is smaller.

Since, *p*-value corresponding to intercept is larger than 0.05, therefore we can not reject the nu hypothesis that the intercept term is zero. Therefore, we may remove the intercept term in t model. This can be done by checking 'constant is zero' checkbox in 'regression' pop up window 'Data Analysis' tool pack in MS Excel.

The *p*-values corresponding to 'Price' and 'AdvEx' are both smaller than 0.05, therefore at 5 level of significance we reject the null hypotheses that 'Price' is insignificant and 'AdvEx' is insignifica in the model.

14.3.6 *F*-Test (ANOVA)

In simple linear regression analysis, *t*-test and *F*-test are used to test the same hypothesis. In case multiple linear regression analysis *F*-test is used to test the overall validity of the model or to test any of the explanatory variables is having linear relationship with the response variable. We test:

$H_0: b_1 = b_2 = \ldots = b_k = 0$ (no explanatory variable is significant)

H_a: a at least one $b_j \neq 0$ (at least one r explanatory variable affects Y linearly).

We proceed in the same manner as we did in case of simple linear regression analysis. We mak the ANOVA table as below:

ANOVA	*d.f*	*Sum of Squares*	*Mean Sum of Squares*	*F-ratio*
Regression or Explained	1	*SSR*	$MSR = SSR$	$F_c = MSR/MSE$
Residual or Unexplained	$N - k = 1$	*SSE*	$MSE = SSE/(n - k = 1)$	
Total	$N - 1$	*SST*		

Under the normality of residuals, the test statistic F_c follows Snedecor's F distribution with , $n-k-1$) d.f. We reject the null hypothesis for this right tailed test when the computed value is igher than the critical value F_α.

The test is illustrated using MS Excel in the following example.

xample 14.11: For the model fitted in Example 14.8, test the overall validity of the model.

olution: We wish to test

$H_0: b_1 = b_2 = ... = b_k = 0$ (no explanatory variable is significant)

H_a: at least one $b_j \neq 0$ (at least one r explanatory variable affects Y linearly)

We use 'Data Analysis' tool pack of MS Excel to fit the regression model. Below, we present the econd table of the Excel outcome:

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.03	14730.01	6.538607	0.012006
Residual	12	27033.31	2252.776		
Total	14	56493.33			

Significance F is the p-value of the test which is less than 0.05. Therefore, at 5% level of ignificance, we reject the null hypothesis and conclude that at least one explanatory variable has ignificant linear relationship with response variable and the fitted linear model is valid.

14.3.7 Confidence Interval Estimates of the Regression Coefficients

he confidence interval estimate of b_j ($j = 1, 2, ..., k$) at a given confidence coefficient $(1-\alpha)$ can be btained by

$$\hat{b}_j \mp t_{\alpha/2}(d.f. = n-k-1) \times \text{Standard Error}(\hat{b}_j)$$

here Standard Error $(\hat{b}_j)$ is given by $\sqrt{MSE \times C_{jj}}$. Here C_{jj} is the $(j, j)^{th}$ element of matrix $X'X)^{-1}$ and MSE is obtained in ANOVA table.

Example 14.12: Construct a 95% confidence interval estimate of regression coefficients for the odel fitted in Example 14.8.

Solution: We use 'Data Analysis' tool pack of MS Excel to fit the regression model. Below, we resent the last table of the Excel outcome partially:

	Coefficients	Standard Error	t-Stat	P-value	Lower 95%	Upper 95%
Intercept	306.5261933	114.2539	2.682851	0.019932	57.58834	555.464
Price	–24.97508952	10.83213	–2.30565	0.039788	–48.5763	–1.37392
Adv. Ex.	74.13095749	25.96732	2.854779	0.014494	17.55303	130.7089

In the above table, last two columns provide the 95% confidence interval for intercept term an other two regression slopes.

14.3.8 Confidence Interval of Mean Response and Prediction Interval of Individual Response

$(1 - \alpha) \times 100\%$ confidence interval estimate of the mean response at some given data-point

$$x_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{pmatrix}$$

is given by

$$\hat{y}_i \mp t_{\alpha/2}(d.f. = n-k-1)\sqrt{MSE \times x_i'(X'X)^{-1}\, x_i}$$

The formula is bit complicated as it involves matrix calculations. Also, confidence interval estimat of mean response cannot be obtained using MS Excel. We suggest using SPSS to obtain confidenc interval. The usage of SPSS is displayed later in this chapter.

$(1-\alpha) \times 100\%$ prediction interval of an individual response for given data-point x_i is obtained a

$$\hat{y}_i \mp t_{\alpha/2}(d.f. = n-k-1)\sqrt{MSE \times (1 + x_i'(X'X)^{-1}\, x_i)}$$

Like confidence interval, prediction interval can also be obtained using SPSS explained later i Section 14.6.

14.4 PROBLEM OF MULTICOLINEARITY

Let us consider the following example;

Example 14.13: Following data was collected for the sales, number of advertisements published and advertizing expenditure for 12 weeks.

Sales ('0,000 Rs.)	*Ads (Nos.)*	*Adv Ex ('000 Rs.)*
43.6	12	13.9
38	11	12
30.1	9	9.3
35.3	7	9.7
46.4	12	12.3
34.2	8	11.4
30.2	6	9.3
40.7	13	14.3
38.5	8	10.2
22.6	6	8.4
37.6	8	11.2
35.2	10	11.1

Fit a regression model to predict the sales. Test the significant of individual explanatory variable. Also test the overall validity of the model.

Solution: We use 'regression' function of 'Data Analysis' tool pack of MS Excel. The outcome is given as below:

Regression Statistics	
Multiple R	0.827051
R Square	0.684014
Adjusted R Square	0.613794
Standard Error	3.988885
Observations	12

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	309.9858	154.9929	9.74112	0.005604
Residual	9	143.2008	15.9112		
Total	11	453.1867			

	Coefficients	*Standard Error*	*t-Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6.583578	8.542155	0.770716	0.460628	−12.7401	25.90728
Ads (Nos.)	0.624675	1.120277	0.557608	0.590709	−1.90957	3.158918
Adv. Ex. ('000 Rs.)	2.138864	1.47015	1.454861	0.17968	−1.18685	5.464574

From the last table, the fitted model is

Predicted Sales = 6.583578 + 0.624675 Ads + 2.138864 AdvEx

From the first table, $R^2 = 0.684014$, adjusted $R^2 = 0.613794$. Thus the model seems to be good

Overall validity of the model (using ANOVA)

We observe that *p*-value (*Significance F*) in ANOVA table is 0.005604 which is much smalle than 0.05. Therefore we reject the null hypothesis and conclude that at least one explanatory variabl is significant.

Test of individual explanatory variable

From the last table, we observe that all three *p*-values are much higher than 0.05. Therefore w accept all the null hypotheses here and conclude that the intercept, Ads, and AdvEx, all three, ar insignificant in the model.

Thus, here we find some contradiction in the conclusions of *F*-test and *t*-test. This is an indicatio of the problem of multicollinearity.

When we regress Y on $X_1, X_2, \ldots, X_k$, we assume that all explanatory variables $X_1, X_2, \ldots, X_k$ are statistically independent of one other. All the explanatory variables affect the values of Y. However. one explanatory variable should not affect the values of other explanatory variable. Sometimes. i practice this assumption is not met and we face the problem of multicollinearity. The correlated variables contribute redundant information to the model and we get some contradictory results o unstable estimates. Some Indications of strong multicollinearity are as follows:

(i) Coefficient signs may not match prior expectations.

(ii) Large change in the value of a previous coefficient when a new variable is added to the model.

(iii) A previously significant variable becomes insignificant when a new explanatory variable is added.

(iv) *F*-test says at least one variable is significant, but none of the *t*'s indicates a useful variable.

(v) Large standard error and corresponding variable is still significant.

(vi) MSE is very high and/or R^2 is very small.

Some other examples where it is likely to face the problem of multicollinearity are given as below:

(i) Miles per gallon of motorized vehicles regressed on horsepower of engine and size of engine.

(ii) Income of a person regressed on age and experience.

4.4.1 Variance Inflationary Factor

ulticollinearity generated by variable X_j is measured by variance inflationary factor given by

$$VIF_j = \frac{1}{1 - R_j^2}$$

here R_j^2 is the coefficient of determination of a linear regression model that uses X_j as the response ariable and all other X variables as the explanatory variables. If $VIF_j > 5$, X_j is highly correlated ith the other explanatory variables.

We below present a partial outcome of regression analysis done using SPSS to show the VIF_j sing the data of Example 14.12.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	6.584	8.542		0.771	0.461		
	No_Adv	0.625	1.12	0.234	0.558	0.591	0.199	5.022
	Ex_Adv	2.139	1.47	0.611	1.455	0.18	0.199	5.022

[a] Dependent Variable: Sales

We can see that both VIF are greater than 5 which indicates the existence of multicollinearity. Tolerance is the reciprocal of VIF (i.e., Tolerance = 1/VIF).

Mathematically, the problem of multicollinearity occurs when the matrix $X'X$ is singular or non-invertible. The least squares estimates cannot be obtained when the matrix $X'X$ is singular. The matrix $X'X$ becomes singular when the columns of matrix X have exact linear dependence. This problem can be identified by obtaining the eigen values of matrix $X'X$. If any of the eigen value of matrix $X'X$ is zero, the matrix $X'X$ is singular. Thus, near-zero eigen value is also an indication of multicollinearity.

Below, we present a partial outcome of SPSS to show the eigen values of matrix $X'X$ for the data of Example 14.12.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	No_Adv	Ex_Adv
1	1	2.966	1	0	0	0
	2	0.03	9.882	0.33	0.17	0
	3	0.003	30.417	0.67	0.83	1

[a] Dependent Variable : Sales

We note that one eigen value is 0.003 which is very close to zero and hence an indication multicollinearity.

The methods of dealing with multicollinearity are – (i) collecting additional data, and (ii) variabl elimination. Collecting additional data may not always be possible. In case it is possible, it may n help in dealing with the problem of mulitcollinearlty always. Below, we describe some methods variable elimination.

14.4.2 Variable Elimination

There are mainly three techniques of variable elimination: (i) stepwise method, (ii) forward inclusio method, and (iii) backward elimination method.

Stepwise Method

This method is based on the *p*-value of *F*-test obtained in ANOVA. The method is illustrated a below:

Suppose we have a response variable and 5 explanatory variables X_1, X_2, X_3, X_4, X_5.

Step 1: Fit 5 simple linear regression models:

- Predicted $Y = a + b_1X_1$
- Predicted $Y = a + b_2X_2$
- Predicted $Y = a + b_3X_3$
- Predicted $Y = a + b_4X_4$
- Predicted $Y = a + b_5X_5$

We obtain the *p*-value of *F*-test for all the five simple linear regression models mentioned above. Suppose that the *p*-value corresponding to the fourth model (having X_4) is the smallest (and below 0.05 as well), then the first explanatory variable included in the model is X_4.

Step 2: Fit 4 two-variable linear regression models having X_4 as one explanatory variable:

- Predicted $Y = a + b_4X_4 + b_1X_1$
- Predicted $Y = a + b_4X_4 + b_2X_2$
- Predicted $Y = a + b_4X_4 + b_3X_3$
- Predicted $Y = a + b_4X_4 + b_5X_5$

We again obtain the *p*-value of *F*-test for all four models of this step. Suppose that the *p*-value corresponding to the third model (having X_4 and X_3) is the smallest (and below 0.05 as well), then the second explanatory variable included in the model is X_3.

Step 3: Fit 3 three-variable linear regression models having X_4 and X_3 as two explanatory variable:

- Predicted $Y = a + b_3X_3 + b_4X_4 + b_1X_1$
- Predicted $Y = a + b_3X_3 + b_4X_4 + b_2X_2$
- Predicted $Y = a + b_3X_3 + b_4X_4 + b_5X_5$

Suppose *p*-value is not below 0.05 for any of these models. We stop here. The best model is the one with X_3 and X_4 only.

orward Inclusion Method

Ve start having only intercept term in the model and no explanatory variable. We include the xplanatory variables one by one as follow:

(i) First, we select the explanatory variable for inclusion in the model which has the highest correlation with Y. Suppose this variable is X_1.

(ii) Now we conduct ANOVA and obtain p-value. If this p-value is smaller than the pre selected entry level (0.05 usually), we include X_1 in the model.

(iii) The second variable selected for inclusion (suppose X_2) in the model is the one which has largest partial correlation with Y. Partial correlation is obtained after adjusting the effect of previous entered variable on Y.

(iv) We again conduct ANOVA for the updated model (having X_1 and X_2 both) and obtain p-value.

(v) If this p-value is smaller than pre selected entry level (0.05), X_2 is included in the model.

(vi) The procedure stops when the p-value at a particular step exceeds pre selected entry level (0.05) or when all the explanatory variables are entered.

Backward Elimination Method

In this method, we start having all the variables in the model. Using the similar criterion discussed in forward inclusion method we keep eliminating the variable at each step. Here we compare the p-values with a pre selected removal level (0.10 usually). If the p-value is larger, corresponding explanatory variable is eliminated from the model.

Because of calculation complication, it is suggested to employ theses methods using some statistical software, like SPSS. We demonstrate the usage of SPSS in section 14.6.

Another important method of dealing with the problem of multicollinearity is ridge regression which is beyond the scope of this book.

14.5 QUALITATIVE EXPLANATORY VARIABLES

When one or more explanatory variables are qualitative, we cannot fit a regression model as regression analysis requires numerical values. In such cases we define dummy variable for the qualitative variable under study. The idea of dummy variable is illustrated through following example:

Example14.14: ABC Inc. provides maintenance service for electronic security systems. In order to estimate the service cost, the managers want to predict the repair time necessary for each maintenance request. Repair time is believed to be related to two factors - number of months since the last maintenance service and type of repair problem (mechanical or electronic). Data for a sample of 10 service calls are given:

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	electronic	2.9
2	6	mechanical	3
3	8	electronic	4.8
4	3	mechanical	1.8
5	2	electronic	2.9
6	7	electronic	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electronic	4.4
10	6	electronic	4.5

Fit the regression model to predict the repair time.

Solution: Let Y denote the repair time, X_1 denote the number of months since last maintenance service.

Using MS Excel, we fit the regression model Y regressed on X_1 as below:

Summary Output

Regression statistics	
Multiple R	0.731
R Square	0.534
Adjusted R Square	0.476
Standard Error	0.781
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	5.596	5.596	9.174	0.016
Residual	8	4.880	0.610		
Total	9	10.476			

	Coefficients	*Standard Error*	*t-Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.147	0.605	3.549	0.008	0.752	3.542
Months Since Last Service	0.304	0.100	3.029	0.016	0.073	0.536

We observe that the using F-test and t-test, the model is valid and X_1 is significantly correlated ith Y.

In order to improve the model, if we wish to estimate the information given in terms of the 'type repair', we need to define a dummy variable as

$X_2 = 0$, if the type of repair is mechanical; 1, if the type of repair is electronic. Using this we pdate the data as below:

Y	X_2	X_1
2	1	2.9
6	0	3
8	1	4.8
3	0	1.8
2	1	2.9
7	1	4.9
9	0	4.2
8	0	4.8
4	1	4.4
6	1	4.5

We again fit the model using MS Excel. The outcome is given below:

Summary Output

Regression statistics	
Multiple R	0.929
R Square	0.863
Adjusted R Square	0.824
Standard Error	1.089
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	52.197	26.098	22.002	0.001
Residual	7	8.303	1.186		
Total	9	60.500			

	Coefficients	*Standard Error*	*t-Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	–1.028	1.332	– 0.772	0.465	– 4.176	2.121
X_2	–3.012	0.736	– 4.094	0.005	– 4.752	– 1.272
X_1	2.182	0.352	6.195	0.000	1.349	3.015

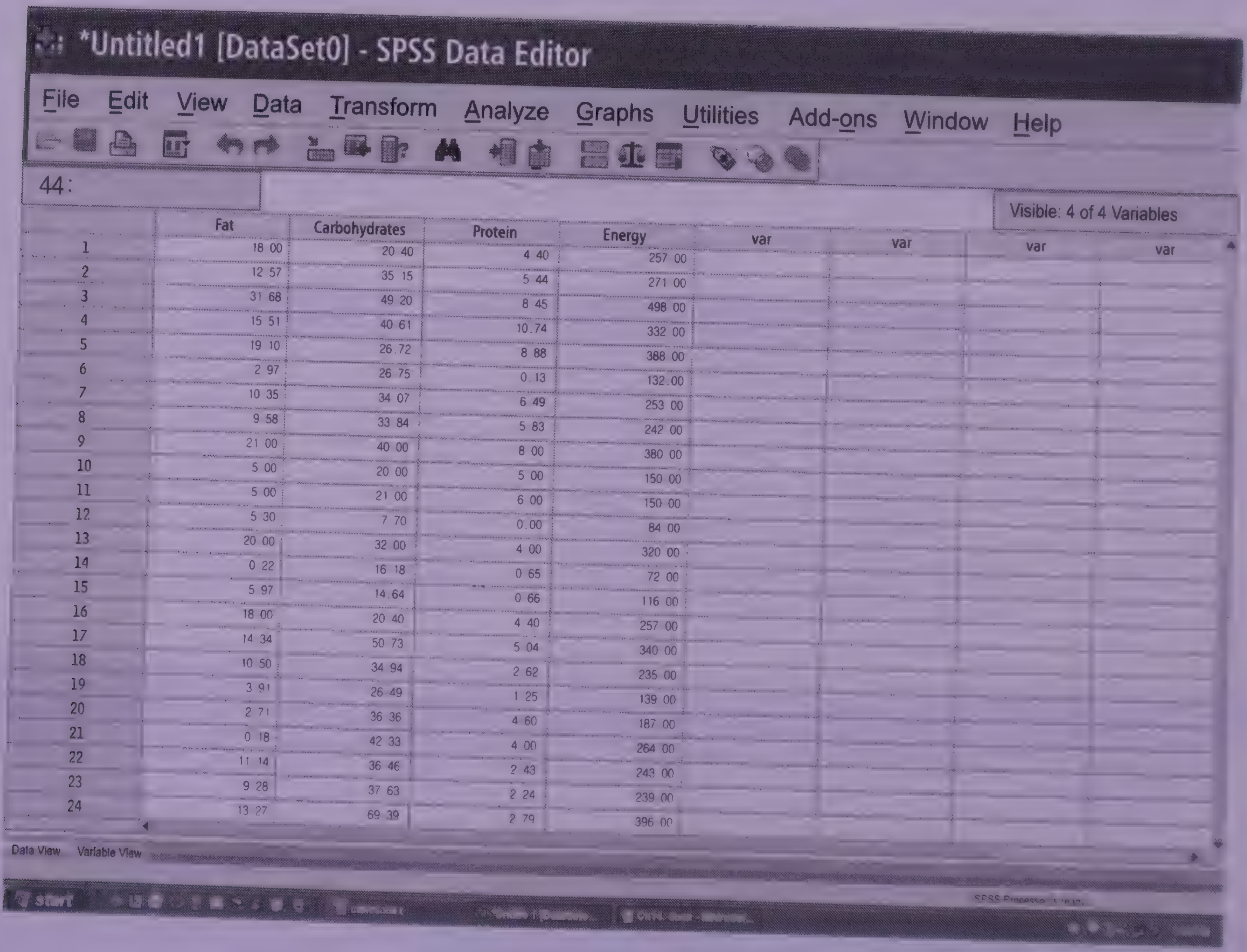
We observe that this model is also valid (F-test) and both explanatory variables are signific (t-test). It can also be observed that the second model is improved (compare adjusted R^2).

14.6 USING SPSS

We illustrate the use of SPSS in regression analysis through the following example:

Example 14.15: In order to fit a model to predict the calories in a cake, 59 cakes were sampled a fats (in grams), carbohydrates (in grams), proteins (in grams), and energy (in kcals) are measure which are given in the file cakes.xls. Fit the regression model and examine the validity of the mod

Solution: Enter the data in SPSS worksheet. You can copy it from MS Excel also. You can chan the names of the variables by clicking 'Variable View' in the left-down corner of the screen. Belo is the screen shot of the worksheet:



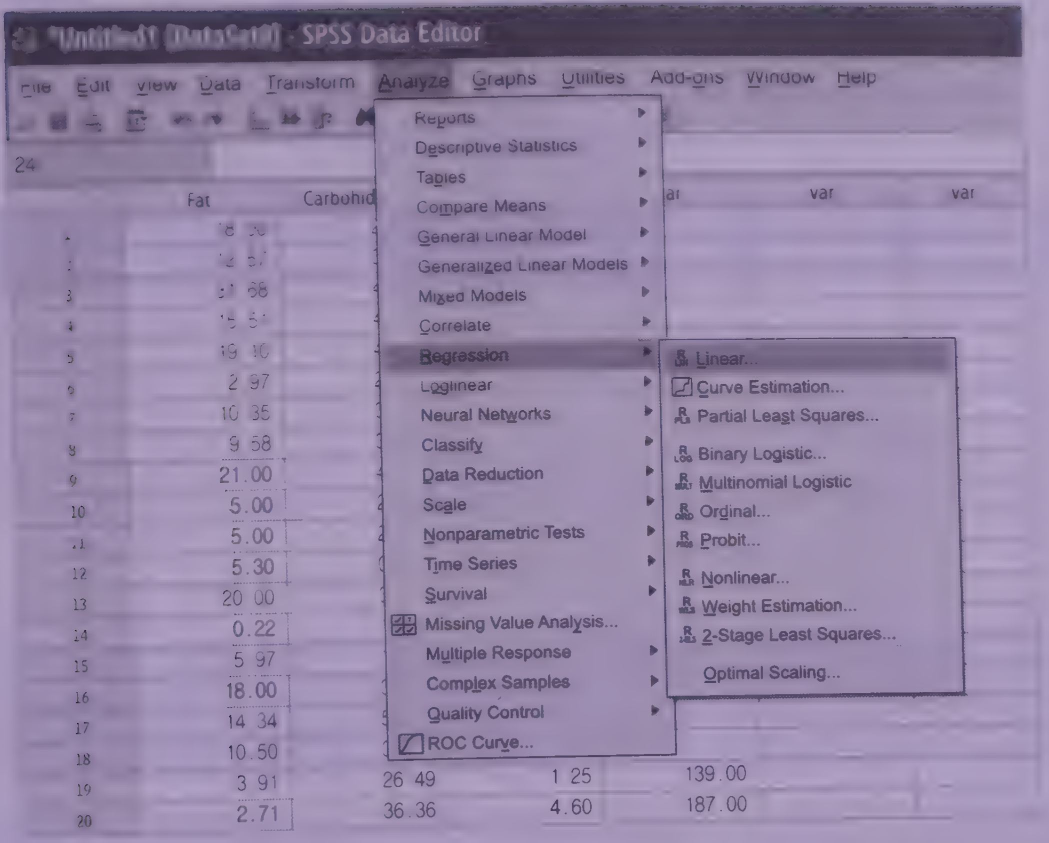
*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
44:
Visible: 4 of 4 Variables
Data View Variable View

	Fat	Carbohydrates	Protein	Energy	var	var	var	var
1	18.00	20.40	4.40	257.00				
2	12.57	35.15	5.44	271.00				
3	31.68	49.20	8.45	498.00				
4	15.51	40.61	10.74	332.00				
5	19.10	26.72	8.88	388.00				
6	2.97	26.75	0.13	132.00				
7	10.35	34.07	6.49	253.00				
8	9.58	33.84	5.83	242.00				
9	21.00	40.00	8.00	380.00				
10	5.00	20.00	5.00	150.00				
11	5.00	21.00	6.00	150.00				
12	5.30	7.70	0.00	84.00				
13	20.00	32.00	4.00	320.00				
14	0.22	16.18	0.65	72.00				
15	5.97	14.64	0.66	116.00				
16	18.00	20.40	4.40	257.00				
17	14.34	50.73	5.04	340.00				
18	10.50	34.94	2.62	235.00				
19	3.91	26.49	1.25	139.00				
20	2.71	36.36	4.60	187.00				
21	0.18	42.33	4.00	264.00				
22	11.14	36.46	2.43	243.00				
23	9.28	37.63	2.24	239.00				
24	13.27	69.39	2.79	396.00				

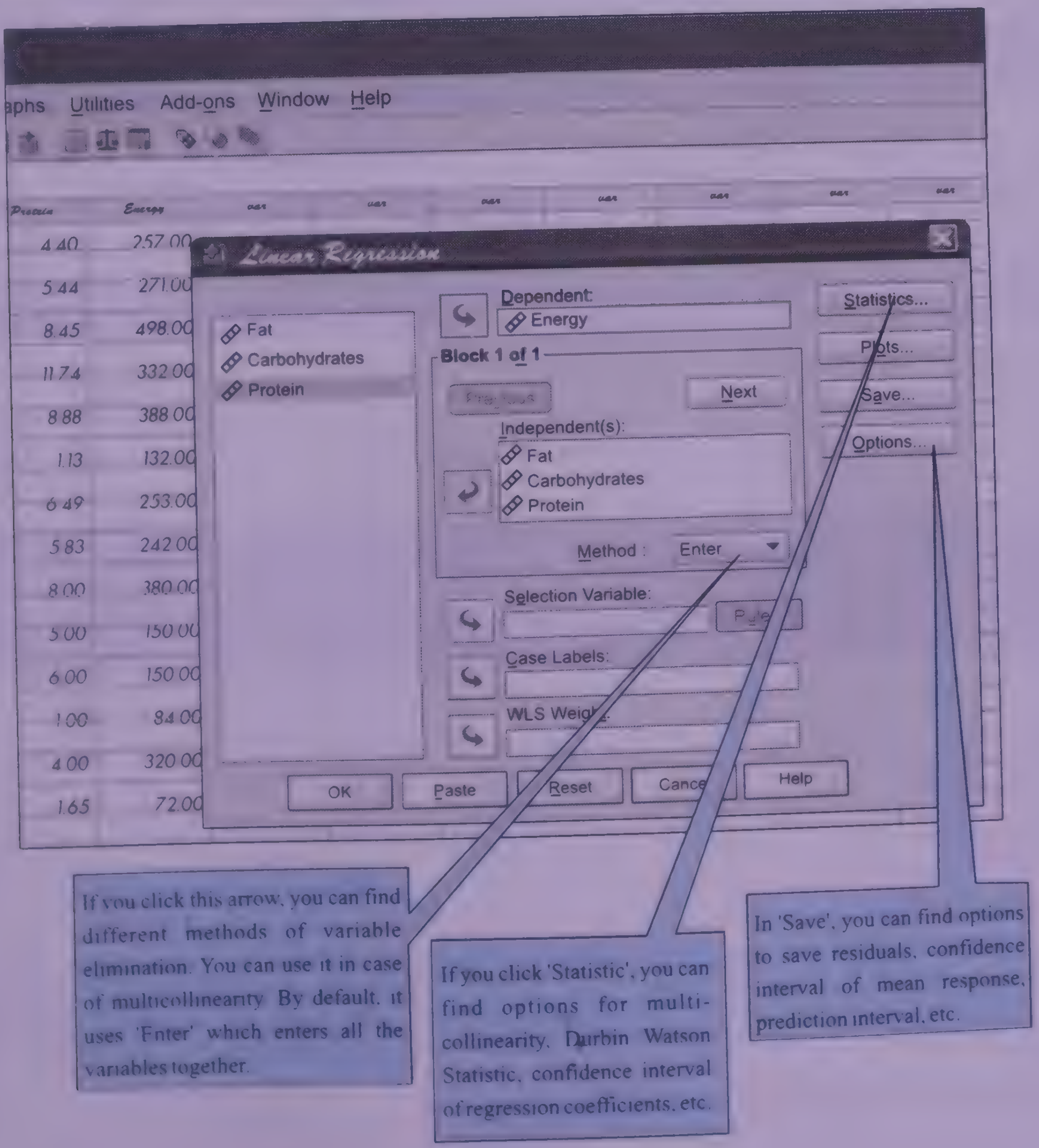
Click 'Analyze'.

Select 'Regression' from the dropdown menu.

Select 'Linear' from the dropdown menu.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
Fat
Carbohid
var
var
var

You will get a popup window. In that window, select 'Energy' as the 'Dependent,' and remaining hree as 'Independent(s)'.


aphs Utilities Add-ons Window Help
Protein
Energy
Linear Regression
Fat
Carbohydrates
Protein
Block 1 of 1
Previous
Next
Selection Variable:
Rule...
Case Labels:
WLS Weight:
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Save...
Options...

Dependent: Energy

Independent(s): Fat, Carbohydrates, Protein

Method : Enter

If you click this arrow, you can find different methods of variable elimination. You can use it in case of multicollinearity. By default, it uses 'Enter' which enters all the variables together.

If you click 'Statistic', you can find options for multicollinearity, Durbin Watson Statistic, confidence interval of regression coefficients, etc.

In 'Save', you can find options to save residuals, confidence interval of mean response, prediction interval, etc.

When you click 'OK', you get the output which is given below:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Protein, Carbohydrates, Fat[a]		Enter

[a] All requested variables entered

[b] Dependent Variable: Energy

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.999[a]	0.998	0.998	4.22125

[a] Predictors: (Constant), Protein, Carbohydrates, Fat

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	602976.637	3	200992.212	1.128E4	0.000[a]
	Residual	980.041	55	17.819		
	Total	603956.678	58			

[a] Predictors: (Constant), Protein, Carbohydrates, Fat

[b] Dependent Variable: Energy

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.239	1.804		2.904	0.005
	Fat	8.802	0.104	0.612	84.581	0.000
	Carbohydrates	3.779	0.047	0.493	79.648	0.000
	Protein	3.709	0.278	0.088	13.320	0.000

[a] Dependent Variable: Energy

If you compare the magnitude of 'Standardized coefficients BETA' you can decide which explanatory variable is more important in explaining Y. Here 'Carbohydrates' is the most important variable in the model.

All other outcomes have already been explained.

ample 14.16: Excel Data file CO2.xlsx contains the data for 123 countries on the following riables:

CO2 emissions (metric tons) per capita (CO2)
GDP per capita PPP (current int. dollar) (GDP)
Electric power consumption (kWh per capita) (Elec)
Population Density (People per sq. km of land area) (Pop_Des)
Inflation on Consumer Prices (annual %) (Inf_CP)
Forest area (% of land area) (F_Area)

t a suitable linear regression model to predict CO2 emission.

lution: See the SPSS outcome below:

Model Summary

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*
1	.842[a]	.709	.697	3.91474

[a] Predictors: (Constant), F_Area, Elec, Pop_Des, Inf_CP, GDP

ANOVA[a]

Model		*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
1	Regression	4375.507	5	875.101	57.102	.000[b]
	Residual	1793.050	117	15.325		
	Total	6168.557	122			

[a] Dependent Variable: CO2

[b] Predictors: (Constant), F_Area, Elec, Pop_Des, Inf_CP, GDP

Coefficients[a]

Model		Unstandardized B	Coefficients std. error	Standardized coefficients Beta	t	Si
1	(Constant)	–.694	1.041		–.666	.507
	GDP	.000	.000	.954	12.964	.000
	Elec	–7.150E-5	.000	–.062	–.927	.356
	Pop_Des	–.002	.001	–.233	–4.426	.000
	Inf_CP	.215	.096	.125	2.231	.028
	F_Area	–.011	.018	–.033	–.642	.522

[a] Dependent Variable: CO2

Variables Entered/Removed[a]

Model	Variables entered	Variables removed	Method
1	F_Area, Elec, Pop_Des, Inf_CP, GDP[b]		Enter

[a] Dependent Variable: CO2

[b] All requested variables entered.

As we can see in the table **Coefficients,** p-values related to the regressors Elec and F_Ar are more than 0.05, we accept the null hypothesis that these regressors are not linearly related wi CO2. Therefore, we discard them from the model and fit the linear model with remaining regressor The SPSS outcome of the same is as below:

Variables Entered/Removed[a]

Model	Variables entered	Variables removed	Method
1	Inf_CP, Pop_Des, GDP[b]		Enter

[a] Dependent Variable: CO2

[b] All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R square	Std. error of the estimate
1	.840[a]	.706	.699	3.90254

[a] Predictors: (Constant), Inf_CP, Pop_Des, GDP

ANOVA[a]

Model	R	Sum of squares	df	Mean square	F	Sig.
1	Regression	4356.208	3	1452.069	95.344	.000[b]
	Residual	1812.349	119	15.230		
	Total	6168.557	122			

[a] Dependent Variable: CO2

[b] Predictors: (Constant), Inf_CP, Pop_Des, GDP

Coefficients[a]

Model		Unstandardized B	Coefficients std. error	Standardized coefficients Beta	t	Sig.
1	(Constant)	–1.141	.798		–1.429	.156
	GDP	.000	.000	.916	16.145	.000
	Pop_Des	–.002	.001	–.221	–4.305	.000
	Inf_CP	.224	.095	.131	2.370	.019

[a] Dependent Variable: CO2

Now we can see that all the p-values in the table **Coefficients** are below 0.05. Thus, each of the egressors is linearly related with CO2. The p-value in ANOVA table (given by Sig.) is below 0.05. herefore, the linear model is overall valid.

The final model is: Predicted CO2 = – 1.141 + 0.000 GDP – 0.002 Pop_Des + 0.224 Inf_CP

The least squares estimate of the regression coefficient corresponding to GDP is coming to e 0.000, but we cannot discard GDP from the model as corresponding p-value is below 0.05, that eans GDP is a significant variable. Also, the value of standardised coefficient of GDP is 0.916, having the highest magnitude. This means GDP is the most important variable in the model. In or to see the effect of GDP clearly, we can transform the GDP values in thousand international doll by dividing by 1000. Thus we fit the model again using SPSS and get the following model:

Predicted CO2 = – 1.141 + 0.329 (GDP/1000) – 0.002 Pop_Des + 0.224 Inf_CP

Rest of the outcomes do not change.

Example 14.17: Excel Data file Life_Ex.xlsx contains data for 89 countries on Life Expectan (Life_Ex)

Air Quality (measured as annual mean PM10 ug/m^3) (AirQ)
Mobile phone subscriptions (per 100 population) (Mob)
Electricity consumption per capita (kWh per person) (Elec)

Fit a suitable linear regression model which can be used to predict life expectancy.

Solution: SPSS outcome of the linear regression analysis is given below:

Variables Entered/Removed[a]

Model	*Variables entered*	*Variables removed*	*Method*
1	Elec, Mob, AirQ[b]		Enter

[a] Dependent Variable: Life_Ex

[b] All requested variables entered.

Model Summary

Model	*R*	*R square*	*Adjusted R square*	*Std. error of the estimate*
1	.682[a]	.466	.447	4.70770

[a] Predictors: (Constant), Elec, Mob, AirQ

ANOVA[a]

Model		*Sum of squares*	*df*	*Mean square*	*F*	*Sig*
1	Regression	1641.427	3	547.142	24.688	.000[b]
	Residual	1883.809	85	22.162		
	Total	3525.236	88			

[a] Dependent Variable: Life_Ex

[b] Predictors: (Constant), Elec, Mob, AirQ

Coefficients[a]

Model		Unstandardized B	Coefficients std. error	Standardized coefficients Beta	t	Sig.
1	(Constant)	72.612	2.128		34.127	.000
	AirQ	–.058	.011	–.441	–5.300	.000
	Mob	.041	.016	.206	2.510	.014
	Elec	.000	.000	.309	3.702	.000

[a] Dependent Variable: Life_Ex

The fitted linear regression model is:

Predicted Life Expectancy = 72.612 – 0.058 AirQ + 0.41 Mob + 0.000 Elec

AirQ is most important regressor as the magnitude of corresponding standardised coefficient is gest. The regressor Elec is the second important variable. In order to see the clear effect of Elec predicted value, we may do the transformation as in the previous example.

ample 14.18: Excel data file UP_Cities_Census.xlsx contains the data for 71 cities in the state Uttar Pradesh on Area, No of Households, Female Population in the age group 0-6 years, and al Literates Population. Fit a suitable linear regression model to predict Female Population in the e group 0-6.

lution: The SPSS outcome is as below:

Variables Entered/Removed[a]

Model	Variables entered	Variables removed	Method
1	Total_Literate_Population, Area, No_Households[b]		Enter

[a] Dependent Variable: Female_Population_0_6

[b] All requested variables entered.

Model Summary[b]

Model	R	R square	Adjusted R square	Std. error of the estimate	Durbin-Watson
1	.981[a]	.963	.961	7544.19972	1.433

[a] Predictors: (Constant), Total_Literate_Population, Area, No_Households

[b] Dependent Variable: Female_Population_0_6

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Si
1	Regression	98423802960.000	3	32807934320.000	576.438	.00
	Residual	3813301608.000	67	56914949.380		
	Total	102237104600.000	70			

[a] Dependent Variable: Female_Population_0_6

[b] Predictors: (Constant), Total_Literate_Population, Area, No_Households

Coefficients[a]

Model	B	Unstandardized coefficients std. error	Standardized coefficients Beta	t	Sig.	Collinearity tolerance	Statist VI
1 (Constant)	5665.257	1479.844		3.828	.000		
Area	5.619	24.829	.013	.226	.822	.164	6.09
No_ Households	.555	.067	1.831	8.275	.000	.011	87.98
Total_ Literate_ Population	–.070	.017	–.872	–4.055	.000	.012	83.10

[a] Dependent Variable: Female_Population_0_6

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition index	Variance Proportions			
				(Constant)	Area	No_ households	Total_ Literat population
1	1	3.502	1.000	.02	.01	.00	.00
	2	.445	2.806	.63	.00	.00	.00
	3	.050	8.386	.35	.98	.01	.02
	4	.003	32.213	.01	.02	.98	.98

[a] Dependent Variable: Female_Population_0_6

In the above outcome, we notice that Area is not significant variable at 5% level of significance. we discard Area and fit the model again. We can notice that VIF values for both of the regressors ne new model are large. This implies the existence of the problem of multicollinearity in the data. rder to deal with this problem, we do variable elimination by using stepwise method. SPSS come of the same is given below:

Variables Entered/Removed[a]

Model	*Variables entered*	*Variables removed*	*Method*
1	No_Households		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	Total_Literate_Population		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

[a] Dependent Variable: Female_Population_0_6

Model Summary[c]

Model	*R*	*R Square*	*Adjusted R square*	*Std. Error of the estimate*	*Durbin-Watson*
1	.976[a]	.954	.953	8298.45466	
2	.981[b]	.963	.962	7491.38422	1.443

[a] Predictors: (Constant), No_Households

[b] Predictors: (Constant), No_Households, Total_Literate_Population

[c] Dependent Variable: Female_Population_0_6

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Si
1	Regression	97485464430.000	1	97485464430.000	1415.616	.00
	Residual	4751640137.000	69	68864349.810		
	Total	102237104600.000	70			
2	Regression	98420887610.000	2	49210443810.000	876.866	.00
	Residual	3816216953.000	68	56120837.550		
	Total	102237104600.000	70			

[a] Dependent Variable: Female_Population_0_6

[b] Predictors: (Constant), No_Households

[c] Predictors: (Constant), No_Households, Total_Literate_Population

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.	Collinearity	Statistic
		B	Std. Error	Beta			tolerance	VIF
1	(Constant)	5699.376	1307.177	4.360	.000			
	No_Households	.296	.008	.976	37.625	.000	1.000	1.000
2	(Constant)	5864.636	1180.741		4.967	.000		
	No_Households	.558	.065	1.843	8.630	.000	.012	83.101
	Total_Literate_Population	-.070	.017	-.872	-4.083	.000	.012	83.101

[a] Dependent Variable: Female_Population_0_6

Excluded Variables[a]

Model	Beta In	t	Sig.	Partial correlation	Collinearity	Statistics	
					tolerance	VIF	Minimu toleran
1 Total_Literate_Population	–.872[b]	–4.083	.000	–.444	.012	83.101	.012

[a] Dependent Variable: Female_Population_0_6

[b] Predictors in the Model: (Constant), No_Households

Collinearity Diagnostics[a]

Model	Dimension	Eigen value	Condition index	Variance Proportions (Constant)	No_Households	Total_Literate_population
1	1	1.658	1.000	.17	.17	
	2	.342	2.200	.83	.83	
2	1	2.553	1.000	.06	.00	.00
	2	.443	2.400	.94	.00	.00
	3	.003	27.299	.00	1.00	1.00

[a] Dependent Variable: Female_Population_0_6

We can see that in the first step, No_Households is included and in the second step tal_Literate_Population is also included in the model. However, the inclusion of al_Literate_Population results in the problem of multicollinearity, which can be realized by high ues of VIF in the second model. This happens because in stepwise method the decision of inclusion a variable is taken using p value of ANOVA which is almost zero in both of the models. In this nario, we will go with the first model. This will avoid the instability in the model crept by lticollinearity. Thus the final model is:

Predicted Female_Population_0_6 = 5699.376 + 0.296 No_Households

PROBLEMS

. Life expectancy at birth may depend on

(i) Adolescent fertility rate (births per 1,000 women ages 15–19)
(ii) Crude Birth rate (per 1,000 people)
(iii) Health expenditure per capita (USD)
(iv) Immunization (% of children ages 12–23 months)
(v) Incidence of tuberculosis (per 100,000 people)
(vi) Infant Mortality rate (per 1,000 live births)
(vii) GDP per capita (USD)

A sample of 181 countries on above mentioned variables is given in LifeEx.xls.

(a) Fit a regression model to predict the life expectancy.
(b) Test the overall validity of the model.
(c) Test for the significance of individual variable.
(d) Obtain R^2 and adjusted R^2 and comment on the values.
(e) Is there any indication of multicollinearity problem, if yes eliminate the variable responsible for this problem and improve the model?
(f) Obtain the prediction interval of individual response and confidence interval of mean response. at each data-point in the sample.

2. Total number of crimes in a state may depend on
   - (i) Violent crime
   - (ii) Police funding
   - (iii) Percentage of 16 to 19 year-olds not in school
   - (iv) Percentage of 20 to 24 year-olds in

   Suppose that the data, provided in file crime.xls, on above variables is collected from 23 states of Ind

   - (a) Fit a regression model to predict the total crimes.
   - (b) Test the overall validity of the model.
   - (c) Test for the significance of individual variable.
   - (d) Obtain R^2 and adjusted R^2 and comment on the values.
   - (e) Is there any indication of multicollinearity problem, if yes eliminate the variable responsible this problem and improve the model?
   - (f) Obtain the prediction interval of individual response and confidence interval of mean response. each data-point in the sample.
3. Explain the difference between R^2 and adjusted R^3.
4. Write a note on the usage of ANOVA in linear regression analysis.
5. Write a note on having intercept in a linear regression model.
6. Why do we use t-test in linear regression analysis.

15

Chapter

actor Analysis

ctor analysis is by far the most often used multivariate technique of research studies, specially taining to social and behavioural sciences. It is a technique applicable when there is a systematic erdependence among a set of observed or manifest variables and the researcher is interested in ding out something more fundamental or latent which creates this commonality. For instance, we ght have data, say, about an individual's income, education, occupation and dwelling area and want nfer from these some factor (such as social class) which summarises the commonality of all the d four variables. The technique used for such purpose is generally described as factor analysis. ctor analysis, thus, seeks to resolve a large set of measured variables in terms of relatively few egories, known as factors. This technique allows the researcher to group variables into factors sed on correlation between variables) and the factors so derived may be treated as new variables ten termed as latent variables) and their value derived by summing the values of the original iables which have been grouped into the factor. The meaning and name of such new variables is jectively determined by the researcher. Since the factors happen to be linear combinations of a, the coordinates of each observation or variable is measured to obtain what are called factor dings. Such factor loadings represent the correlation between the particular variable and the tor, and are usually place in a matrix of correlations between the variable and the factors.

5.1 THE MATHEMATICAL BASIS

e mathematical basis of factor analysis concerns a data matrix (also termed as score matrix). mbolized as *S*. The matrix contains the scores of *N* persons of *k* measures. Thus a_1 is the score of rson 1 on measure a, a_2 is the score of person 2 on measure a, and k_N is the score of person N on asure k. The score matrix then take the form as shown following:

SCORE MATRIX (or Matrix *S*)

Measures (variables)

		a	*b*	*c*	*k*
	1	a_1	b_1	c_1	k_1
	2	a_2	b_2	c_2	k_2
	3	a_3	b_3	c_3	k_3
Persons (objects)	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	N	a_N	b_N	c_N	k_N

It is assumed that scores on each measure are standardized [i.e., $x_i = (X - \bar{X}_i)^2/\sigma_i$] . T being so, the sum of scores in any column of the matrix, *S*, is zero and the variance of scores in column is 1.0. Then factors (a factor is any linear combination of the variables in a data matrix can be stated in a general way like: $A = W_a a + W_b b + \ldots + W_k k$) are obtained (by any method factoring). After this, we work out factor loadings (i.e., factor-variable correlations). Then communal symbolized as h^2, the eigen value and the total sum of squares are obtained and the results interpret For realistic results, we resort to the technique of rotation, because such rotations reveal differ structures in the data. Finally, factor scores are obtained which help in explaining what the fact mean. They also facilitate comparison among groups of items as groups. With factor scores, one also perform several other multivariate analyses such as multiple regression, cluster analysis, multi discriminant analysis, etc.

15.2 IMPORTANT METHODS OF FACTOR ANALYSIS

There are several methods of factor analysis, but they do not necessarily give same results. As su factor analysis is not a single unique method but a set of techniques. Important methods of fac analysis are:

(i) the centroid method;
(ii) the principal components method;
(iii) the maximum likelihood method.

Before we describe these different methods of factor analysis, it seems appropriate that so basic terms relating to factor analysis be well understood.

(i) *Factor:* A factor is an underlying dimension that account for several observed variabl There can be one or more factors, depending upon the nature of the study and the numb of variables involved in it.

(ii) *Factor-loadings*: Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor-variable correlations. In fact, factor-loadings work as key to understanding what the factors mean. It is the absolute size (rather than the signs, plus or minus) of the loadings that is important in the interpretation of a factor.

(iii) *Communality* (h^2): Communality, symbolized as h^2, shows how much of each variable is accounted for the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:

$$h^2 \text{ of the } i\text{th variable} = (i\text{th factor loading of factor } A)^2 + (i\text{th factor loading of factor } B)^2 + \ldots$$

(iv) *Eigen value (or latent root)*: When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

(v) *Total sum of squares*: When eigen values of all factors are totalled, the resulting value is termed as the total sum of squares. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for what all the variables taken together represent. If the variables are all very different from each other, this index will be low. If they fall into one or more highly redundant groups, and if the extracted factors account for all the groups, the index will then approach unity.

(vi) *Rotation*: Rotation, in the context of factor analysis, is something like staining a microscope slide. Just as different stains on it reveal different structures in the tissue, different rotations reveal different structures in the data. Though different rotations give results that appear to be entirely different, but from a statistical point of view, all results are taken as equal, none superior or inferior to others. However, from the standpoint of making sense of the results of factor analysis, one must select the right rotation. If the factors are independent orthogonal rotation is done and if the factors are correlated, an oblique rotation is made. Communality for each variables will remain undisturbed regardless of rotation but the eigen values will change as result of rotation.

(vii) *Factor scores*: Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factors mean. With such scores, several other multivariate analyses can be performed.

We can now take up the important methods of factor analysis.

5.2.1 Centroid Method

his method of factor analysis, developed by L.L. Thurstone, was quite frequently used until about 950 before the advent of large capacity high speed computers. The centroid method tends to maximize e sum of loadings, disregarding signs; it is the method which extracts the largest sum of absolute loadings for each factor in turn. It is defined by linear combinations in which all weights are eith +1.0 or –1.0. The main merit of this method is that it is relatively simple, can be easily understood a involves simpler computations. If one understands this method, it becomes easy to understand t mechanics involved in other methods of factor analysis.

Various steps involved in this method are as follows:

(i) This method starts with the computation of a matrix of correlations, *R*, wherein unities a place in the diagonal spaces. The product moment formula is used for working out t correlation coefficients.

(ii) If the correlation matrix so obtained happens to be positive manifold (i.e., disregarding t diagonal elements each variable has a large sum of positive correlations than of negati correlations), the centroid method requires that the weights for all variables be +1.0. other words, the variables are not weighted; they are simply summed. But in case t correlation matrix is not a positive manifold, then reflections must be made before the fi centroid factor is obtained.

(iii) The first centroid factor is determined as under:

(a) The sum of the coefficients (including the diagonal unity) in each column of the correlati matrix is worked out.

(b) Then the sum of these column sums (*T*) is obtained.

(c) The sum of each column obtained as per (a) above is divided by the square root of obtained in (b) above, resulting in what are called centroid loadings. This way eac centroid loading (one loading for one variable) is computed. The full set of loadings s obtained constitute the first centroid factor (say *A*).

(iv) To obtain second centroid factor (say *B*), one must first obtain a matrix of residu coefficients. For this purpose, the loadings for the two variables on the first centroid fact are multiplied. This is done for all possible pairs of variables (in each diagonal space is th square of the particular factor loading). The resulting matrix of factor cross products ma be named as Q_1. Then Q_1 is subtracted element by element from the original matrix c correlation, *R*, and the result is the first matrix of residual coefficients, R_1.

One should understand the nature of the elements in R_1 matrix. Each diagonal element is a parti variance i.e., the variance that remains after the influence of the first factor is partialed. Each of diagonal element is a partial co-variance i.e., the covariance between two variables after the influenc of the first factor is removed. This can be verified by looking at the partial correlation coefficien between any two variables say 1 and 2 when factor *A* is held constant

$$r_{12\cdot A} = \frac{r_{12} - r_{1A} \cdot r_{2A}}{\sqrt{1 - r_{1A}^2}\sqrt{1 - r_{2A}^2}}$$

(The numerator in the above formula is what is found in R_1 corresponding to the entry for riables 1 and 2. In the denominator, the square of the term on the left is exactly what is found in the agonal element for variable 1 in R_1. Likewise the partial variance for 2 is found in the diagonal ace for that variable in the residual matrix.)

Since in R_1 the diagonal terms are partial variances and the off-diagonal terms are partial ovariances, it is easy to convert the entire table to a matrix of partial correlations. For this purpose ne has to divide the elements in each row by the square-root of the diagonal element for that row d then dividing the elements in each column by the square-root of the diagonal element for that lumn.

After obtaining R_1, one must *reflect* some of the variables in it, meaning thereby that some of the riables are given negative signs in the sum [This is usually done by inspection. The aim in doing this ould be to obtain a reflected matrix, R'_1, which will have the highest possible sum of coefficients)]. For any variable which is so reflected, the signs of all coefficients in that column and row of the sidual matrix are changed. When this is done, the matrix is named as 'reflected matrix' form which e loadings are obtained in the usual way (already explained in the context of first centroid factor), t the loadings of the variables which were reflected must be given negative signs. The full set of adings so obtained constitutes the second centroid factor (say B). Thus loadings on the second entroid factor are obtained from R'_1.

(v) For subsequent factors (C, D, etc.) the same process outlined above is repeated. After the second centroid factor is obtained, cross products are computed forming, matrix, Q_2. This is then subtracted from R_1 (and not from R'_1) resulting in R_2. To obtain a third factor (C), one should operate on R_2 in the same way as on R_1. First, some of the variables would have to be reflected to maximize the sum of loadings, which would produce R'_2. Loadings would be computed from R'_2 as they were from R'_1. Again, it would be necessary to give negative signs to the loadings of variables which were reflected which would result in third centroid factor (C).

We may now illustrate this method by an example.

xample 15.1: Given is the following correlation matrix, R, relating to eight variables with unities in the iagonal spaces:

Variables

		1	2	3	4	5	6	7	8
	1	1.000	.709	.204	.081	.626	.113	.155	.774
	2	.709	1.000	.051	.089	.581	.098	.083	.652
	3	.204	.051	1.000	.671	.123	.689	.582	.072
Variables	4	.081	.089	.671	1.000	.022	.798	.613	.111
	5	.626	.581	.123	.022	1.000	.047	.201	.724
	6	.113	.098	.689	.798	.047	1.000	.801	.120
	7	.155	.083	.582	.613	.201	.801	1.000	.152
	8	.774	.652	.072	.111	.724	.120	.152	1.000

Using the centroid method of factor analysis, work out the first and second centroid factors fr the above information.

Solution: Given correlation matrix, R, is a positive manifold and as such the weights for all variab be +1.0. Accordingly, we calculate the first centroid factor (A) as under:

Table 15.1(a)

Variables

		1	2	3	4	5	6	7	8
	1	1.000	.709	.204	.081	.626	.113	.155	.774
	2	0.709	1.000	.051	.089	.581	.098	.083	.652
	3	0.204	.051	1.000	.671	.123	.689	.582	.072
Variables	4	0.081	.089	.671	1.000	.022	.798	.613	.111
	5	0.626	.581	.123	.022	1.000	.047	.201	.724
	6	0.113	.098	.689	.798	.047	1.000	.801	.120
	7	0.155	.083	.582	.613	.201	.801	1.000	.152
	8	0.774	.652	.072	.111	.724	.120	.152	1.000
Column sums		3.662	3.263	3.392	3.385	3.324	3.666	3.587	3.605
Sum of the column sums (T) = 27.884 $\therefore \sqrt{T} = 5.281$									
First centroid factor $A = \frac{3.662}{5.281}, \frac{3.263}{5.281}, \frac{3.392}{5.281}, \frac{3.385}{5.281}, \frac{3.324}{5.281}, \frac{3.666}{5.281}, \frac{3.587}{5.281}, \frac{3.605}{5.281}$ = .693, .618, .642, .641, .629, .694, .679, .683									

We can also state this information as under:

Table 15.1 (b)

Variables	*Factor loadings concerning first centroid factor A*
1	0.693
2	0.618
3	0.642
4	0.641
5	0.629
6	0.694
7	0.679
8	0.683

To obtain the second centroid factor B, we first of all develop (as shown on the next page) th first matrix of factor cross product, Q_1:

First Matrix of Factor Cross Product (Q_1)

First centroid factor *A*	0.693	0.618	0.642	0.641	0.629	0.694	0.679	0.683
0.693	0.480	0.428	0.445	0.444	0.436	0.481	0.471	0.473
0.618	0.428	0.382	0.397	0.396	0.389	0.429	0.420	0.422
0.642	0.445	0.397	0.412	0.412	0.404	0.446	0.436	0.438
0.641	0.444	0.396	0.412	0.411	0.403	0.445	0.435	0.438
0.629	0.436	0.389	0.404	0.403	0.396	0.437	0.427	0.430
0.694	0.481	0.429	0.446	0.445	0.437	0.482	0.471	0.474
0.679	0.471	0.420	0.436	0.435	0.427	0.471	0.461	0.464
0.683	0.473	0.422	0.438	0.438	0.430	0.474	0.464	0.466

Now we obtain first matrix of residual coefficient (R_1) by subtracting Q_1 from R as shown low:

First Matrix of Residual Coefficiént (R_1)
Variables

		1	2	3	4	5	6	7	8
	1	0.520	0.281	−0.241	−0.363	0.190	−0.368	−0.316	0.301
	2	0.281	0.618	−.346	−0.307	0.192	−0.331	−0.337	0.230
	3	−0.241	−0.346	0.588	0.259	−0.281	0.243	0.146	−0.366
Variables	4	−0.363	−0.307	0.259	0.589	−0.381	0.353	0.178	−0.327
	5	0.190	0.192	−0.281	−0.381	0.604	−0.390	−0.217	0.294
	6	−0.368	−0.331	0.243	0.353	−0.390	0.518	0.330	−0.354
	7	−0.316	−0.337	0.146	0.178	−0.226	0.330	0.539	−0.312
	8	0.301	0.230	−0.366	−0.327	0.294	−0.354	−0.312	0.534

Reflecting the variables 3, 4, 6 and 7, we obtain reflected matrix of residual coefficient (R'_1) as nder and then we can extract the second centroid factor (B) from it as shown on the next page.

Reflected Matrix of Residual Coefficients (R'_1)
and Extraction of 2nd Centroid Factor (B)

Variables

		1	2	3*	4*	5	6*	7*	8
	1	0.520	0.281	0.241	0.363	0.190	0.368	0.316	0.301
	2	0.281	0.618	0.346	0.307	0.192	0.331	0.337	0.230
	3*	0.241	0.346	0.588	0.259	0.281	0.243	0.146	0.366
Variables	4*	0.363	0.307	0.259	0.589	0.381	0.353	0.178	0.327
	5	0.190	0.192	0.281	0.381	0.604	0.390	0.217	0.294
	6*	0.368	0.331	0.243	0.353	0.390	0.518	0.330	0.354
	7*	0.316	0.337	0.146	0.178	0.226	0.330	0.539	0.312
	8	0.301	0.230	0.366	0.327	0.294	0.354	0.312	0.534
Column sums:		2.580	2.642	2.470	2.757	2.558	2.887	2.375	2.718

Sum of column sums (T) = 20.987 $\therefore$ $\sqrt{T} = 4.581$

Second centroid factor B = .563 .577 –.539 –.602 .558 –.630 –.518 .593

*These variables were reflected.

Now we can write the matrix of factor loadings as under:

Variables	*Factor loadings*	
	Centroid Factor A	*Centroid Factor B*
1	0.693	0.563
2	0.618	0.577
3	0.642	–0.539
4	0.641	–0.602
5	0.629	0.558
6	0.694	–0.630
7	0.679	–0.518
8	0.683	0.593

Example 15.2: Work out the communality and eigen values from the final results obtained in Examp 15.1. Also explain what they (along with the said two factors) indicate.

Solution: We work out the communality and eigen values for the given problem as under:

Table 15.2

Variables	*Factor loadings*		*Communality (h^2)*
	Centroid factor A	*Centroid factor B*	
1	.693	.563	$(.693)^2+(.563)^2=.797$
2	.618	.577	$(.618)^2+(.577)^2=.715$
3	.642	–.539	$(.642)^2+(-.539)^2=.703$
4	.641	–.602	$(.641)^2+(-.602)^2=.773$
5	.629	.558	$(.629)^2+(.558)^2=.707$
6	.694	–.630	$(.694)^2+(-.630)^2=.879$
7	.679	–.518	$(.679)^2+(-.518)^2=.729$
8	.683	.593	$(.683)^2+(.593)^2=.818$
Eigen value (Variance accounted for i.e., common variance)	3.490	2.631	6.121
Proportion of total variance	.44 (44%)	.33 (33%)	.77 (77%)
Proportion of common variance	.57 (57%)	.43 (43%)	1.00 (100%)

Each communality in the above table represents the proportion of variance in the corresponding ow) variable and is accounted for by the two factors (*A* and *B*). For instance, 79.7% of the variance variable one is accounted for by the centroid factor *A* and *B* and the remaining 20.3% of the total riance in variable one scores is thought of as being made up of two parts: a factor specific to the ribute represented by variable one, and a portion due to errors of measurement involved in the sessment of variable one (but there is no mention of these portions in the above table because we ually concentrate on common variance in factor analysis).

It has become customary in factor analysis literature for a loading of 0.33 to be the minimum solute value to be interpreted. The portion of a variable's variance accounted for by this minimum ading is approximately 10%. This criterion, though arbitrary, is being used more or less by way of nvention, and as such must be kept in view when one reads and interprets the multivariate research sults. In our example, factor *A* has loading in excess of 0.33 on all variables; such a factor is usually lled "*the general factor*" and is taken to represent whatever it is that all of the variables have in mmon. We might consider all the eight variables to be product of some unobserved variable (which n be named subjectively by the researcher considering the nature of his study). The factor name is osen in such a way that it conveys what it is that all variables that correlate with it (that "load on ") have in common. Factor *B* in our example has all loadings in excess of 0.33, but half of them are ith negative signs. Such a factor is called a "*bipolar factor*" and is taken to represent a single mension with two poles. Each of these poles is defined by a cluster of variables—one pole by those ith positive loadings and the other pole with negative loadings.

We can give different names to the said two groups to help us interpret and name factor *B*. The ws at the bottom of the above table give us further information about the usefulness of the two ctors in explaining the relations among the eight variables. The total variance (*V*) in the analysis is ken as equal to the number of variables involved (on the presumption that variables are standardized). this present example, then $V = 8.0$. The row labeled "Eigen value" or "Common variance" gives e numerical value of that portion of the variance attributed to the factor in the concerning column oove it. These are found by summing up the squared values of the corresponding factor loadings. hus the total value, 8.0, is partitioned into 3.490 as eigen value for factor *A* and 2.631 as eigen value or factor *B* and the total 6.121 as the sum of eigen values for these two factors. The corresponding oportion of the total variance, 8.0, are shown in the next row; there we can notice that 77% of the tal variance is related to these two factors, i.e., approximately 77% of the total variance is common ariance whereas remaining 23% of it is made up of portions unique to individual variables and the chniques used to measure them. The last row shows that of the common variance approximately 7% is accounted for by factor *A* and the other 43% by factor *B*. Thus it can be concluded that the vo factors together "explain" the common variance.

5.2.2 Principal Components Method

rincipal components (PC) analysis is a procedure to convert a set of observations of possibly orrelated variables into a set of values of linearly uncorrelated variables called principal components. he number of principal components is less than or equal to the number of original variables. This ansformation is defined in such a way that the first principal component accounts for the largest ariability in the data, and each succeeding component in turn has the highest variance possible under e constraint that it is uncorrelated with the preceding components. Principal components are uaranteed to be independent under some conditions.

The PC technique is very popularly used for factor analysis. PC method of factor analysis se to maximize the sum of squared loadings of each factor extracted in turn. Accordingly, PC fact account for the larger variability in the data than that is accounted for using any other factori method.

The aim of the principal components method is the construction out of a given set of variab X_j's ($j = 1, 2, \ldots, k$) of new variables (p_i), called principal components which are linear combinati of the X_s

$$\begin{aligned} p_1 &= a_{11} X_1 + a_{12} X_2 + \ldots + a_{1k} X_k \\ p_2 &= a_{21} X_1 + a_{22} X_2 + \ldots + a_{2k} X_k \\ &\quad . \qquad\quad . \qquad\quad . \\ &\quad . \qquad\quad . \qquad\quad . \\ p_k &= a_{k1} X_1 + a_{k2} X_2 + \ldots + a_{kk} X_k \end{aligned}$$

The method is being applied mostly by using standardized variables, i.e., $z_j = \left(X_j - \overline{X}_j\right)^2 / \sigma_j$.

The a_{ij}'s are called loadings and are worked out in such a way that the extracted princip components satisfy two conditions: (i) principal components are uncorrelated (orthogonal) and (ii) t first principal component (p_1) has the maximum variance, the second principal component (p_2) h the next maximum variance and so on.

Following steps are usually involved in principal components method

(i) Estimates of a_{ij}'s are obtained with which X's are transformed into orthogonal variabl i.e., the principal components. A decision is also taken with regard to the question: ho many of the components to retain into the analysis?

(ii) We then proceed with the regression of Y on these principal components i.e.,

$$Y = \hat{y}_1 p_1 + \hat{y}_2 p_2 + \ldots + \hat{y}_m p_m \ (m < k)$$

(iii) From the $\hat{a}_{ij}$ and $\hat{y}_{ij}$, we may find b_{ij} of the original model, transferring back from the p into the standardized X's.

Alternative method for finding the factor loadings is as under:

(i) Correlation coefficients (by the product moment method) between the pairs of k variabl are worked out and may be arranged in the form of a correlation matrix, R, as under:

Correlation Matrix, R

		Variables				
		X_1	X_2	X_3	.	X_k
	X_1	r_{11}	r_{12}	r_{13}	....	r_{1k}
	X_2	r_{21}	r_{22}	r_{23}	....	r_{3k}
Variables	X_3	r_{31}	r_{32}	r_{33}	....	r_{3k}
	.	.	.	.		
	.	.	.	.		
	X_k	r_{k1}	r_{k2}	r_{k3}	....	r_{kk}

The main diagonal spaces include unities since such elements are self-correlations. The correlation matrix happens to be a symmetrical matrix.

(ii) Presuming the correlation matrix to be positive manifold (if this is not so, then reflections as mentioned in case of centroid method must be made), the first step is to obtain the sum of coefficients in each column, including the diagonal element. The vector of column sums is referred to as U_{a1} and when U_{a1} is normalized, we call it V_{a1}. This is done by squaring and summing the column sums in U_{a1} and then dividing each element in U_{a1} by the square root of the sum of squares (which may be termed as normalizing factor). Then elements in V_{a1} are accumulatively multiplied by the first row of R to obtain the first element in a new vector U_{a2}. For instance, in multiplying V_{a1} by the first row of R, the first element in V_{a1} would be multiplied by the r_{11} value and this would be added to the product of the second element in V_{a1} multiplied by the r_{12} value, which would be added to the product of third element in V_{a1} multiplied by the r_{13} value, and so on for all the corresponding elements in V_{a1} and the first row of R. To obtain the second element of U_{a2}, the same process would be repeated i.e., the elements in V_{a1} are accumulatively multiplied by the 2nd row of R. The same process would be repeated for each row of R and the result would be a new vector U_{a2}. Then U_{a2} would be normalized to obtain V_{a2}. One would then compare V_{a1} and V_{a2}. If they are nearly identical, then convergence is said to have occurred (If convergence does not occur, one should go on using these trial vectors again and again till convergence occurs). Suppose the convergence occurs when we work out V_{a8} in which case V_{a7} will be taken as V_a (the characteristic vector) which can be converted into loadings on the first principal component when we multiply the said vector (i.e., each element of V_a) by the square root of the number we obtain for normalizing U_{a8}.

(iii) To obtain factor B, one seeks solutions for V_b, and the actual factor loadings for second component factor, B. The same procedures are used as we had adopted for finding the first factor, except that one operates off the first residual matrix, R_1 rather than the original correlation matrix R (We operate on R_1 in just the same way as we did in case of centroid method stated earlier).

(iv) This very procedure is repeated over and over again to obtain the successive PC factors (viz. C, D, etc.).

Other steps involved in factor analysis

(a) Next the question is: How many principal components to retain in a particular study? Various criteria for this purpose have been suggested, but one often used is Kaiser's criterion. According to this criterion only the principal components, having latent root greater than one, are considered as essential and should be retained.

(b) The principal components so extracted and retained are then rotated from their beginning position to enhance the interpretability of the factors.

(c) Communality, symbolized, h^2, is then worked out which shows how much of each variable is accounted for by the underlying factors taken together. A high communality figure means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:

$$h^2 \text{ of the } i\text{th variable} = (i\text{th factor loading of factor } A)^2 + (i\text{th factor loading of factor } B)^2 + \ldots$$

Then follows the task of interpretation. The amount of variance explained (sum of squar loadings) by each PC factor is equal to the corresponding characteristic root. When the roots are divided by the number of variables, they show the characteristic roots as proportio of total variance explained.

(d) The variables are then regressed against each factor loading and the resulting regressi coefficients are used to generate what are known as factor scores which are then used further analysis and can also be used as inputs in several other multivariate analyses.

Example 15.3: Consider the eight variables of Example 15.1 and compute:

(i) The first two principal component factors;

(ii) The communality for each variable on the basis of said two component factors;

(iii) The proportion of total variance as well as the proportion of common variance explained b each of the two component factors.

Solution: Since the given correlation matrix is a positive manifold, we work out the first principa component factor (using trial vectors) as under:

Table 15.3

	Variables							
	1	2	3	4	5	6	7	8
1	1.000	.709	.204	.081	.626	.113	.155	.774
2	.709	1.000	.051	.089	.581	.098	.083	.652
3	.204	.051	1.000	.671	.123	.689	.582	.072
4	.081	.089	.671	1.000	.022	.798	.613	.111
Variables 5	.626	.581	.123	.022	1.000	.047	.201	.724
6	.113	.098	.689	.798	.047	1.000	.801	.120
7	.155	.083	.582	.613	.201	.801	1.000	.152
8	.774	.652	.072	.111	.724	.120	.152	1.000
Column sums U_{a1}	3.662	3.263	3.392	3.385	3.324	3.666	3.587	3.605
Normalizing U_{a1} we obtain V_{a1} i.e., V_{a1} = U_a/Normalizing factor*	.371	.331	.344	.343	.337	.372	.363	.365

*Normalizing factor $= \sqrt{(3.662)^2 + (3.263)^2 + (3.392)^2 + (3.385)^2 + (3.324)^2 + (3.666)^2 + (3.587)^2 + (3.605)^2}$

$= \sqrt{97.372} = 9.868$

Then we obtain U_{a2} by accumulatively multiplying V_{a1} row by row into R and the result comes as nder:

$$U_{a2} : [1.296, 1.143, 1.201, 1.201, 1.165, 1.308, 1.280, 1.275]$$

Normalizing it we obtain (normalizing factor for U_{a2} will be worked out as above and will be 3.493):

$$V_{a2} : [.371, .327, .344, .344, .334, .374, .366, .365]$$

Comparing V_{a1} and V_{a2}, we find the two vectors are almost equal and this shows convergence as occurred. Hence V_{a1} is taken as the characteristic vector, V_a. Finally, we compute the loadings on he first principal component by multiplying V_a by the square root of the number that we obtain for ormalizing U_{a2}. The result is as under:

Variables	(Characteristic vector V_a)	×	$\sqrt{\text{normalizing factor of } U_{a_2}}$	=	Principal Component I
1	.371	×	1.868	=	.69
2	.331	×	1.868	=	.62
3	.344	×	1.868	=	.64
4	.343	×	1.868	=	.64
5	.337	×	1.868	=	.63
6	.372	×	1.868	=	.70
7	.363	×	1.868	=	.68
8	.365	×	1.868	=	.68

For finding principal component II, we have to proceed on similar lines (as stated in the context of obtaining centroid factor B earlier in this chapter) to obtain the following result*:

Variables	Principal component II
1	+.57
2	+.59
3	–.52
4	–.59
5	+.57
6	–.61
7	–.49
8	–.61

The other parts of the question can now be worked out (after first putting the above information in a matrix form) as given below:

---

*This can easily be worked out. Actual working has been left as an exercise for the students.

Variables	*Principal components*		*Communality, h^2*
	I	*II*	
1	0.69	+0.57	$(0.69)^2 + (0.57)^2 = 0.801$
2	0.62	+0.59	$(0.62)^2 + (0.59)^2 = 0.733$
3	0.64	–0.52	$(0.64)^2 + (-0.52)^2 = 0.680$
4	0.64	–0.59	$(0.64)^2 + (-0.59)^2 = 0.758$
5	0.63	+0.57	$(0.63)^2 + (0.57)^2 = 0.722$
6	0.70	–0.61	$(0.70)^2 + (-0.61)^2 = 0.862$
7	0.68	–0.49	$(0.68)^2 + (-0.49)^2 = 0.703$
8	0.68	–0.61	$(0.68)^2 + (-0.61)^2 = 0.835$
Eigen value i.e., common variance	3.4914	2.6007	6.0921
Proportion.436 of total(43.6%) variance	.325 (32.5%)	.761 (76%)	
Proportion.573 of common variance	.427 (57%)	1.000 (43%)	(100%)

All these values can be interpreted in the same manner as stated earlier.

15.2.3 Maximum Likelihood Method

The maximum likelihood (ML) method consists in obtaining sets of factor loadings successively in such a way that each, in turn, explains as much as possible of the population correlation matrix as estimated from the sample correlation matrix. If R_s stands for the correlation matrix actually obtained from the data in a sample, R_p stands for the correlation matrix that would be obtained if the entire population were tested, then the ML method seeks to extrapolate what is known from R_s in the best possible way to estimate R_p (but the PC method only maximizes the variance explained in R_s). Thus, the ML method is a statistical approach in which one maximizes some relationship between the sample of data and the population from which the sample was drawn.

The arithmetic underlying the ML method is relatively difficult in comparison to that involved in the PC method and as such is understandable when one has adequate grounding in calculus, higher algebra and matrix algebra in particular. Iterative approach is employed in ML method also to find each factor, but the iterative procedures have proved much more difficult than what we find in the case of PC method. Hence the ML method is generally not used for factor analysis in practice.

The loadings obtained on the first factor are employed in the usual way to obtain a matrix of the residual coefficients. A significance test is then applied to indicate whether it would be reasonable to extract a second factor. This goes on repeatedly in search of one factor after another. One stops actoring after the significance test fails to reject the null hypothesis for the residual matrix. The final roduct is a matrix of factor loadings. The ML factor loadings can be interpreted in a similar fashion s we have explained in case of the centroid or the PC method.

15.3 ROTATION IN FACTOR ANALYSIS

One often talks about the rotated solutions in the context of factor analysis. This is done (i.e., a factor matrix is subjected to rotation) to attain what is technically called "simple structure" in data. Rotation onstitutes the geometric aspects of factor analysis. Only the axes of the graph (wherein the points epresenting variables have been shown) are rotated keeping the location of these points relative to ach other understood. Simple structure is obtained by rotating the axes until:

(i) Each row of the factor matrix has one zero.

(ii) Each column of the factor matrix has p zeros, where p is the number of factors.

(iii) For each pair of factors, there are several variables for which the loading on one is virtually zero and the loading on the other is substantial.

(iv) If there are many factors, then for each pair of factors there are many variables for which both loadings are zero.

(v) For every pair of factors, the number of variables with non-vanishing loadings on both of them is small.

All these criteria simply imply that the factor analysis should reduce the complexity of all the variables.

15.4 R-TYPE AND Q-TYPE FACTOR ANALYSIS

Factor analysis may be R-type factor analysis or it may be Q-type factor analysis. In *R-type factor analysis*, high correlations occur when respondents who score high on variable 1 also score high on variable 2 and respondents who score low on variable 1 also score low on variable 2. Factors emerge when there are high correlations within groups of variables. In *Q-type factor analysis,* the correlations are computed between pairs of respondents instead of pairs of variables. High correlations occur when respondent 1's pattern of responses on all the variables is much like respondent 2's pattern of responses. Factors emerge when there are high correlations within groups of people. *Q*-type analysis is useful when the object is to sort out people into groups based on their simultaneous responses to all the variables.

Factor analysis has been mainly used in developing psychological tests (such as *IQ* tests, personality tests, and the like) in the realm of psychology. In marketing, this technique has been used to look at media readership profiles of people.

15.5 MERITS AND DEMERITS OF FACTOR ANALYSIS

The main merits of factor analysis can be stated thus:

(i) The technique of factor analysis is quite useful when we want to condense and simplify the multivariate data.

(ii) The technique is helpful in pointing out important and interesting, relationships amo observed data that were there all the time, but not easy to see from the data alone.

(iii) The technique can reveal the latent factors (i.e., underlying factors not directly observe that determine relationships among several variables concerning a research study. F example, if people are asked to rate different cold drinks (say, Limca, Nova-cola, Gold Sp and so on) according to preference, a factor analysis may reveal some salient characteristi of cold drinks that underlie the relative preferences.

(iv) The technique may be used in the context of empirical clustering of products, media people i.e., for providing a classification scheme when data scored on various rating scal have to be grouped together.

One should also be aware of several limitations of factor analysis; important ones are as follow

(i) Factor analysis, like all multivariate techniques, involves laborious computations involvin heavy cost burden. With computer facility available these days, there is no doubt that facto analysis has become relatively faster and easier, but the cost factor continues to be th same i.e., large factor analyses are still bound to be quite expensive.

(ii) The results of a single factor analysis are considered generally less reliable and dependabl for very often a factor analysis starts with a set of imperfect data. "The factors are nothin but blurred averages, difficult to be identified." To overcome this difficulty, it has bee realised that analysis should at least be done twice. If we get more or less similar result from all rounds of analyses, our confidence concerning such results increases.

(iii) Factor-analysis is a complicated decision tool that can be used only when one has thoroug knowledge and enough experience of handling this tool. Even then, at times it may not wor well and may even disappoint the user.

Thus, factor analysis helps the investigator make sense of large bodies of intertwined data. I also points out some interesting relationships that might not have been obvious from examination o the input data alone.

15.6 USING SPSS

We illustrate the use of SPSS in solving factor analysis problem through below example:

Example 15.4: A market researcher wants to determine the underlying benefits consumers seek from the purchase of a car. A sample of 15 respondents was interviewed. The respondents were asked to indicate their degree of agreement with the following statements using a 10-point scale (0 = strongly disagree, 9 = strongly agree).

X_1: I like a car that has stylish interior;

X_2: I prefer a car that looks great;

X_3: I prefer a car giving high mileage;

X_4: I prefer a car with low maintenance;

X_5: I prefer a car that provides a good value for money.

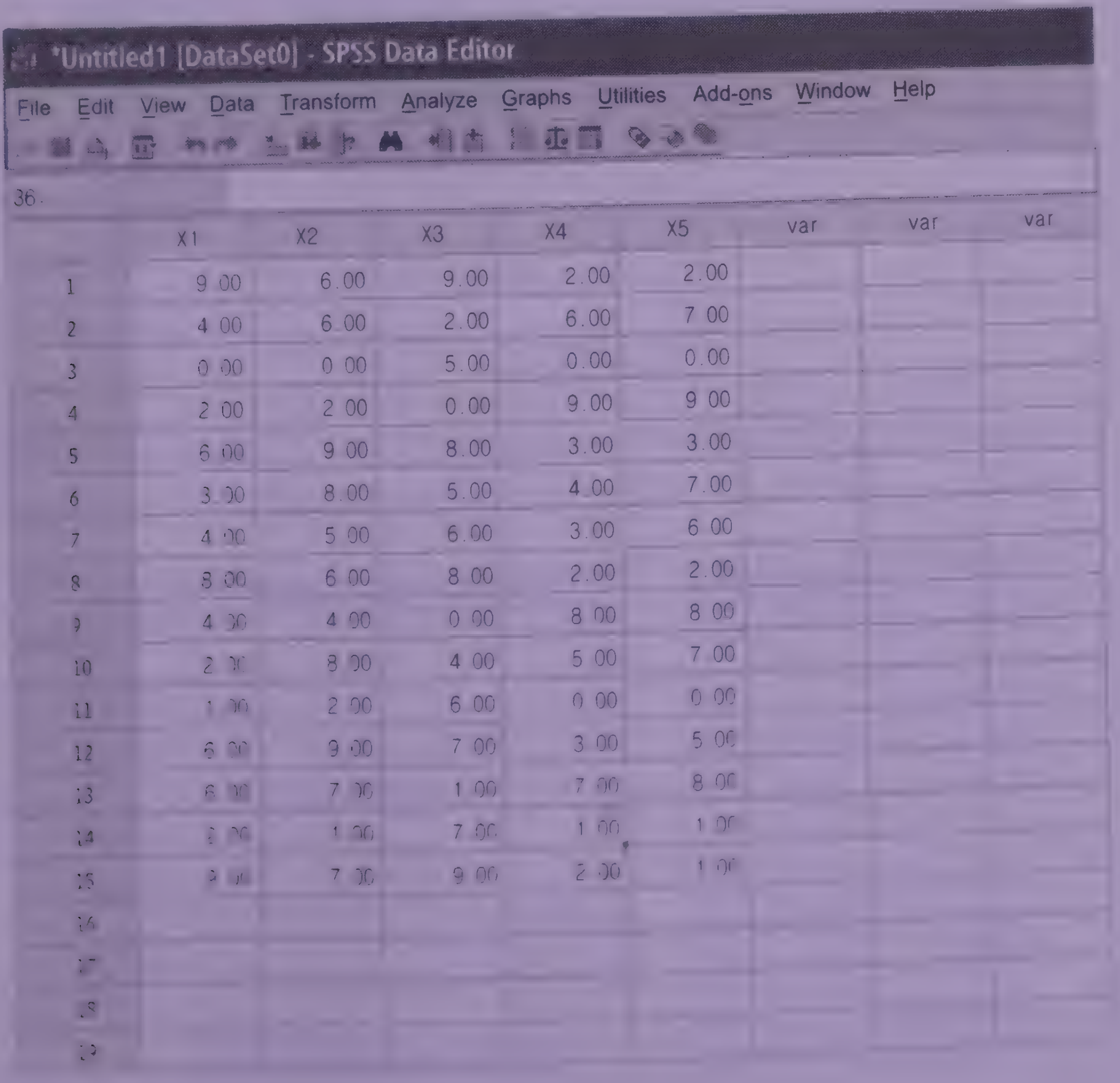
ollected data is given below:

Respondent No.	X_1	X_2	X_3	X_4	X_5
1	9	6	9	2	2
2	4	6	2	6	7
3	0	0	5	0	0
4	2	2	0	9	9
5	6	9	8	3	3
6	3	8	5	4	7
7	4	5	6	3	6
8	8	6	8	2	2
9	4	4	0	8	8
10	2	8	4	5	7
11	1	2	6	0	0
12	6	9	7	3	5
13	6	7	1	7	8
14	2	1	7	1	1
15	9	7	9	2	1

Perform a factor analysis and identify the underlying factors.

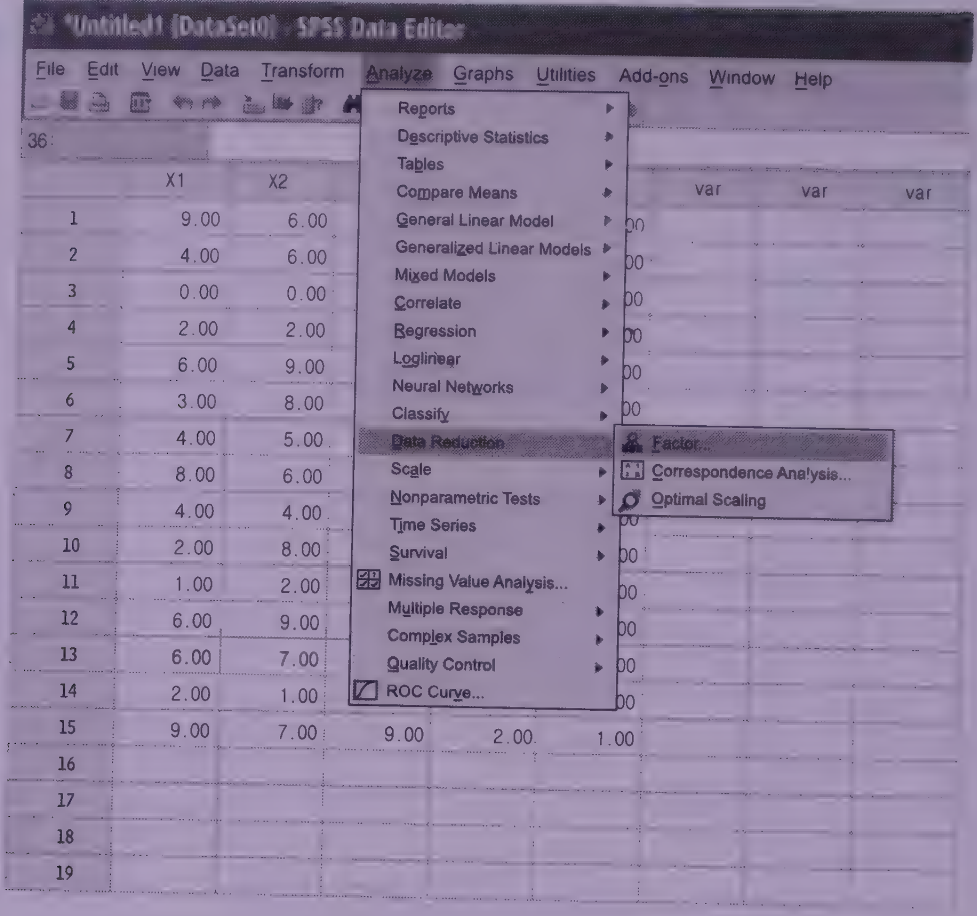
olution: We use SPSS to perform the factor analysis.

We first enter the data in the SPSS worksheet.


*Untitled1 [DataSet0] - SPSS Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
36

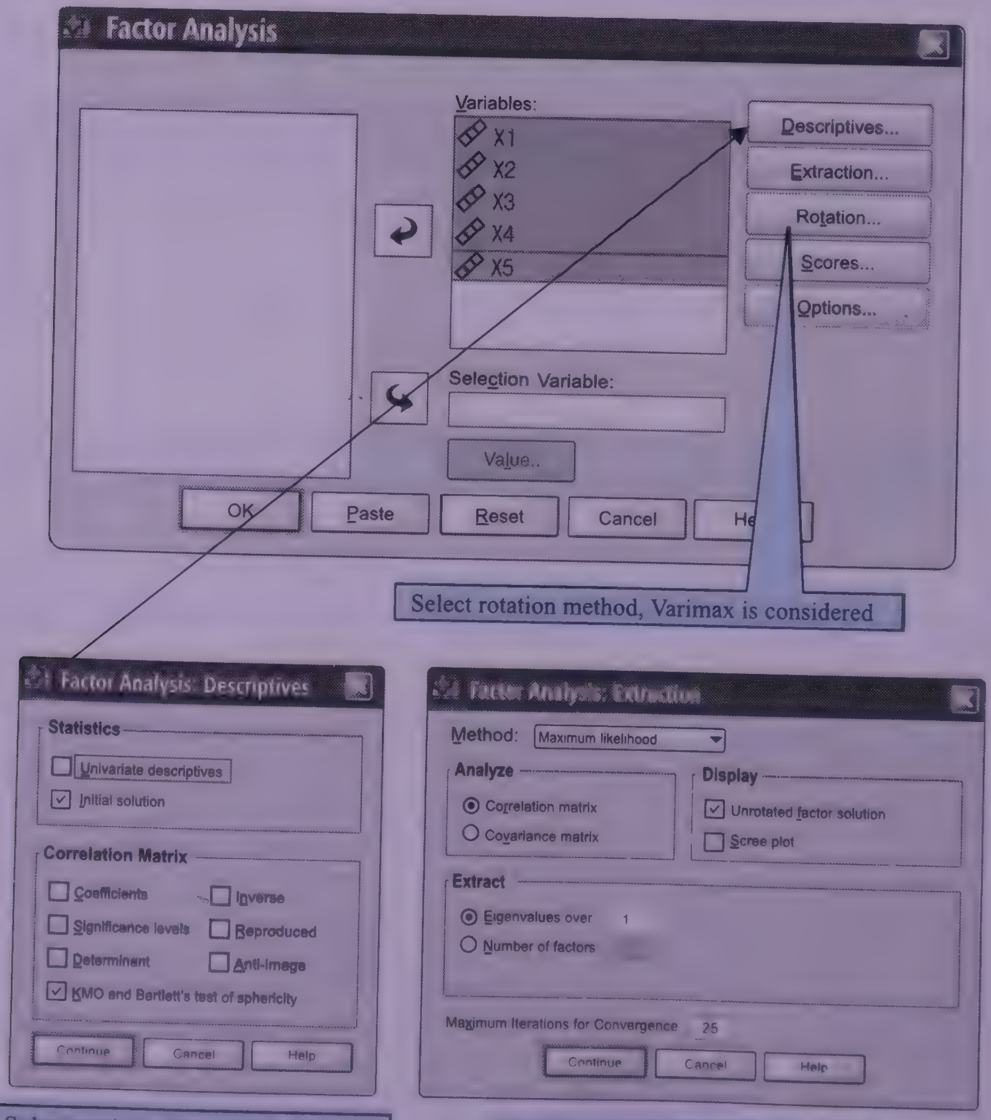
	X1	X2	X3	X4	X5	var	var	var
1	9.00	6.00	9.00	2.00	2.00			
2	4.00	6.00	2.00	6.00	7.00			
3	0.00	0.00	5.00	0.00	0.00			
4	2.00	2.00	0.00	9.00	9.00			
5	6.00	9.00	8.00	3.00	3.00			
6	3.00	8.00	5.00	4.00	7.00			
7	4.00	5.00	6.00	3.00	6.00			
8	8.00	6.00	8.00	2.00	2.00			
9	4.00	4.00	0.00	8.00	8.00			
10	2.00	8.00	4.00	5.00	7.00			
11	1.00	2.00	6.00	0.00	0.00			
12	6.00	9.00	7.00	3.00	5.00			
13	6.00	7.00	1.00	7.00	8.00			
14	2.00	1.00	7.00	1.00	1.00			
15	9.00	7.00	9.00	2.00	1.00			
16								
17								
18								
19								

Click 'Analyze', select 'Data Reduction', select 'Factor'.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
Factor...
Correspondence Analysis...
Optimal Scaling
var
var
var

	X1	X2			
1	9.00	6.00			
2	4.00	6.00			
3	0.00	0.00			
4	2.00	2.00			
5	6.00	9.00			
6	3.00	8.00			
7	4.00	5.00			
8	8.00	6.00			
9	4.00	4.00			
10	2.00	8.00			
11	1.00	2.00			
12	6.00	9.00			
13	6.00	7.00			
14	2.00	1.00			
15	9.00	7.00	9.00	2.00	1.00
16					
17					
18					
19					

Select all the variables as shown below:


Factor Analysis
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Selection Variable:
Value..
OK
Paste
Reset
Cancel
Help
Factor Analysis: Descriptives
Continue
Cancel
Help
Factor Analysis: Extraction
Continue
Cancel
Help

Variables:
X1
X2
X3
X4
X5

Select rotation method, Varimax is considered

Statistics
☐ Univariate descriptives
☑ Initial solution

Correlation Matrix
☐ Coefficients ☐ Inverse
☐ Significance levels ☐ Reproduced
☐ Determinant ☐ Anti-image
☑ KMO and Bartlett's test of sphericity

Method: Maximum likelihood

Analyze
● Correlation matrix
○ Covariance matrix

Display
☑ Unrotated factor solution
☐ Scree plot

Extract
● Eigenvalues over 1
○ Number of factors

Maximum Iterations for Convergence 25

Select various descriptive statistics you wish to see in outcome.

Select the method (PC, ML, etc.)

Outcome is presented and explained below:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.516
Bartlett's Test of Sphericity	Approx. Chi-Square	65.331
	df	10
	Sig.	0.000

KMO Test: Test the suitability of Factor Analysis. This measure varies between 0 and 1, and alues closer to 1 are better.

Bartlett Test of Sphericity: Statistical test for overall significance of the correlations within a orrelation matrix. Uses Chi Square distribution with $p(p-1)/2$ d.f., where p is number of variables. ig. gives the p-value which is .000, less than 0.05 here. Thus there is significant correlation among ariables.

Communalities

	Initial	*Extraction*
X_1	1.000	0.815
X_2	1.000	0.849
X_3	1.000	0.956
X_4	1.000	0.954
X_5	1.000	0.956

Extraction Method: Principal Component Analysis.

Communalities: This is the proportion of each variable's variance that can be explained by the actors. It is also noted as h^2 and can be defined as the sum of squared factor loadings for the ariables. The variance of each variable is standardized to unity and partitioned into two parts — ommunality of that variable and specific variance of that variable, i.e., communality + specific ariance = 1. Therefore, no communality can be more than 1. Communality is due to the correlation mong variables. We explain communality only as specific variance is beyond the control of our and. We used principal component method for extracting communalities. In this method, initial ommunalities are taken as unity.

Total Variance Explained

Component	*Initial Eigen values*			*Extraction sums of squared Loadings*			*Rotation sums of squared Loadings*		
	Total	% of variance	Cumulative %	Total	% of variance	Cumula-tive %	Total	% of variance	Cumula tive %
1	2.755	55.092	55.092	2.755	55.092	55.092	2.736	54.711	54.711
2	1.775	35.497	90.589	1.775	35.497	90.589	1.794	35.878	90.589
3	0.377	7.542	98.131						
4	0.065	1.299	99.431						
5	0.028	0.569	100.000						

Extraction Method: Principal Component Analysis.

Total Variance Explained: We used principal component (PC) method of factor analysis. In PC method factor is called as component. The initial number of factors is the same as the number o variables used in the factor analysis. However, not all 5 factors will be retained. You can choos number of factors or eigenvalue method. In the later method, the number of factors = number o eigenvalues of correlation matrix more than 1. Initial Eigenvalues are the eigenvalues of correlatio matrix. We can see here, only two eigenvalues are more than 1. Also two factors explain 90.589% variance, while 3 factors explain 98.131% variance.

Extraction Sums of Squared Loadings: The number of rows in this panel of the table correspond to the number of factors retained. The values in this panel of the table are calculated in the same way as the values in the left panel. In some other method, these values may be smaller.

Rotation Sums of Squared Loadings: The matrix of factor loadings is rotated orthogonally using Varimax rotation. Total amount of variance accounted for is redistributed over the two extracted factors. This helps making the factors distinct. In our example, total 90.589% variance is redistributed over the 2 factors.

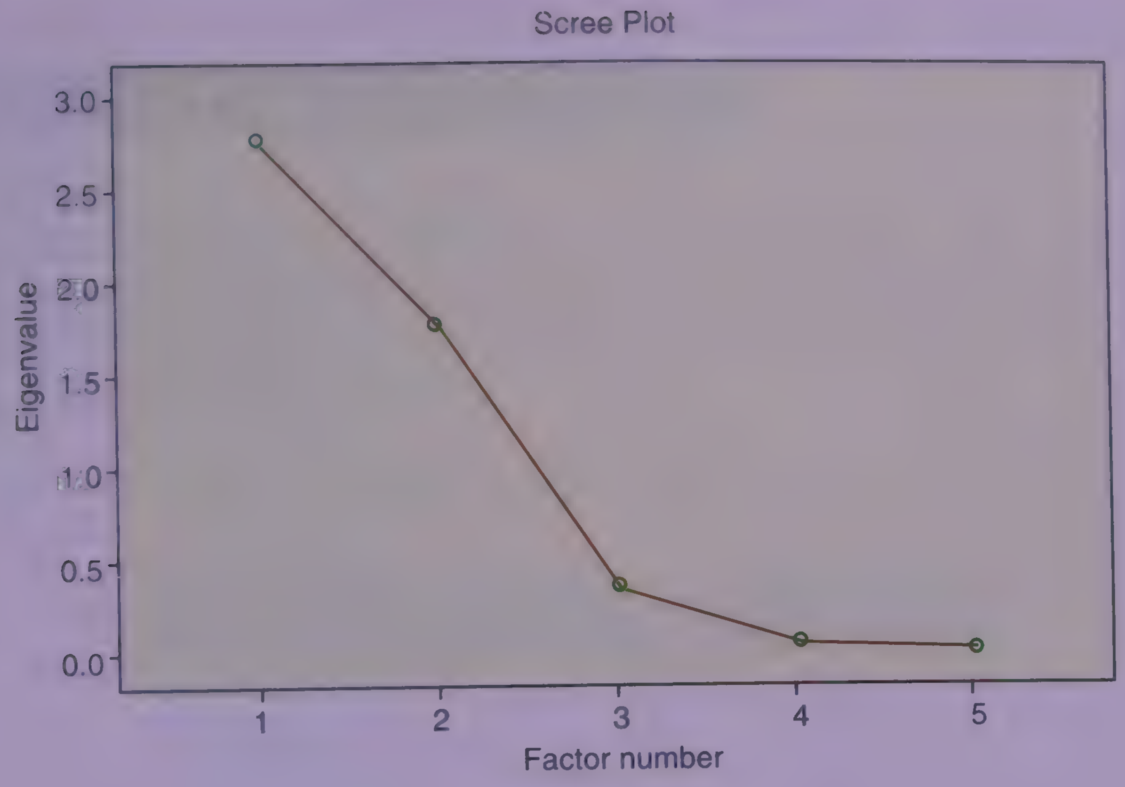


Fig. 15.1

Scree Plot: This is a plot between the eigen value and the factor number. From the third factor on, ou can see that the line is almost flat, meaning the each successive factor is accounting for smaller nd smaller amounts of the total variance. This plot is called a ‘Scree’ Plot because it often looks like ‘scree’ slope, where rocks have fallen down and accumulated on the side of a mountain.

Component Matrix[a]

	Component	
	1	2
X_1	–0.295	0.853
X_2	0.048	0.920
X_3	–0.938	0.278
X_4	0.950	0.228
X_5	0.940	0.268

Extraction Method: Principal Component Analysis

[a] 2 components extracted.

Component Matrix (Factor Matrix): This table contains the unrotated factor loadings, which are the correlations between the variable and the factor.

Reproduced Correlations

		X_1	X_2	X_3	X_4	X_5
Reproduced Correlation	X_1	0.815[a]	0.771	0.514	–0.086	–0.049
	X_2	0.771	0.849[a]	0.211	0.255	0.291
	X_3	0.514	0.211	0.956[a]	–0.827	–0.807
	X_4	–0.086	0.255	–0.827	0.954[a]	0.954
	X_5	–0.049	0.291	–0.807	0.954	0.956[a]
Residual[b]	X_1		–0.161	–0.045	0.068	–0.047
	X_2	–0.161		0.020	–0.065	0.028
	X_3	–0.045	0.020		–0.005	0.033
	X_4	0.068	–0.065	–0.005		–0.027
	X_5	–0.047	0.028	0.033	–0.027	

Extraction Method: Principal Component Analysis.

[a] Reproduced communalities

[b] Residuals are computed between observed and reproduced correlations. There are 3 (30.0%) nonredundant residuals with absolute values greater than 0.05.

Reproduced Correlation: The reproduced correlation matrix is the correlation matrix based on the extracted factors. We want the values in the reproduced matrix to be as close to the values in the original correlation matrix as possible. This means that the residual matrix, which contains the differences between the original and the reproduced matrix to be close to zero. Diagonal elements of the reproduced correlation matrix are extracted communalities.

Rotated Component Matrix[a]

	Component	
	1	2
X_1	–0.173	0.886
X_2	0.175	0.904
X_3	–0.890	0.405
X_4	0.972	0.093
X_5	0.968	0.134

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normal ization

[a] Rotation converged in 3 iterations.

Rotated Component Matrix (Rotated Factor Matrix): This table contains the rotated factor oadings, which are the correlations between the variable and the factor. Clearly,

Factor $f_1 > X_3, X_4, X_5$

Factor $f_2 > X_1, X_2$

Suppose rotated factor matrix is denoted as ^. Factor scores are the observations on factors and an be used for further analysis. Obtain factor score coefficient matrix R^{-1}^. Calculate factor scores or each variable as ZR^{1}^, where R is the correlation matrix and Z is standardized X matrix. Matrix represents the observations on the variables.

Component Transformation Matrix

Component	1	2
1	0.990	– 0.139
2	0.139	0.990

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Transformation Matrix (Factor Transformation Matrix): This is the matrix y which we multiply the unrotated factor matrix to get the rotated factor matrix.

Component Score Coefficient Matrix

	Component	
	1	2
X_1	–0.039	0.491
X_2	0.089	0.511
X_3	–0.315	0.202
X_4	0.359	0.079
X_5	0.359	0.102

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component (Factor) Score Coefficient Matrix: This matrix is used to obtain factor scor The scores of each observation on the underlying factors can be used to represent each factor. Th is common to use these factors as independent variables in regression analysis to reduce t multicollinearity.

Using the factor score coefficients and the observations on the original values, we can obtain t values on factors f_1 and f_2. Factor scores can also be saved as new columns (variables) in SPS The factor scores for our example are given below:

f_1	f_2
−1.26	9.63
4.42	6.60
−1.60	1.00
6.58	3.62
0.17	9.67
2.96	7.57
1.61	6.55
−0.90	8.94
5.96	5.44
3.68	6.96
−1.78	2.71
1.21	9.67
5.47	8.07
−1.51	3.07
−1.53	10.04

Perceptual Map: A scatter plot between the values of f_1 and f_2 given below:

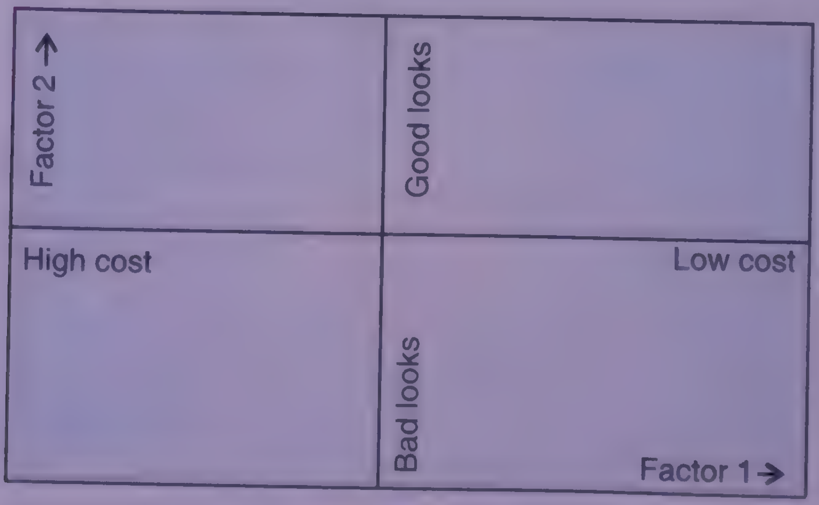


Fig. 15.2

Recall that

- X_1 : I like a car that has stylish interior;
- X_2 : I prefer a car that looks great;
- X_3 : I prefer a car giving high mileage;
- X_4 : I prefer a car with low maintenance;
- X_5 : I prefer a car that provides a good value for money.

Since Factor 1 loads on X_3, X_4, and X_5, it may be called as 'cost factor'. Factor 2 loads on X_1 and $_2$, it may be called as 'looks factor'. The perceptual map helps us studying the perceptions of ustomers or potential customers in four quadrants. For example, first quadrant represent where oth of the factors are preferred, i.e., low cost and good looks.

PROBLEMS

1. What is the main objective of factor analysis?
2. Explain various methods of factor analysis in brief.
3. What is a perceptual map?
4. How can the problem of multicollinearity in regression analysis be solved using factor analysis?
5. What are factor scores?
6. What is a scree plot and how do you interpret it?
7. What is Bartlett test of sphericity in factor analysis?
8. What is rotation in factor analysis?

Discriminant Analysis

16

Chapter

16.1 INTRODUCTION

Discriminant analysis is technique to discriminate between two or more mutually exclusive an exhaustive groups on the basis of some explanatory variables. These groups are known a-priori When the criterion variable has two categories, the technique is known as two-group discriminan analysis. When three or more categories are involved, the technique is referred to as multipl discriminant analysis.

Discriminant analysis helps the researchers who are interested to understand how consumer differ with respect to demographic and psychographic characteristics. Discriminant analysis is als used to predict the group membership. Banks use discriminant analysis to discriminate between the customers who default and who repay the loan in time, based on their age, income, assets, number o dependents, and previous outstanding loan, etc.

Consider the following example:

Example 16.1: Suppose a company wants to determine if one of its new product—a new and improved digital camera will be commercially successful or not. The company wants to distinguish the purchasers from non-purchasers. The company makes a panel of 10 potential purchasers and devised the rating scales on three characteristics—durability, performance, and style. The panel is asked to give the digital camera ratings from 1 to 10 on these three scales. After the product was evaluated, they are asked their buying intention ("would purchase" or "would not purchase"). Five said that they would purchase the digital camera and the other five said that they would not purchase it. The ratings given by 10 panelists are as below:

Groups based on purchase intention	Durability (X_1)	Performance (X_2)	Style (X_3)
Group A: Would purchase			
Person 1	8	9	6
Person 2	6	7	5
Person 3	10	6	3
Person 4	9	4	4
Person 5	4	8	2
Group Mean	**7.4**	**6.8**	**4**
Group B: Would not purchase			
Person 6	5	4	7
Person 7	3	7	2
Person 8	4	5	5
Person 9	2	4	3
Person 10	2	2	2
Group Mean	**3.2**	**4.4**	**3.8**
Difference between Group Means	**4.2**	**2.4**	**0.2**

Let us write the observations on the variable X_1 as below:

	Purchasers					Non-purchasers				
Person	1	2	3	4	5	6	7	8	9	10
X_1	8	6	10	9	4	5	3	4	2	2

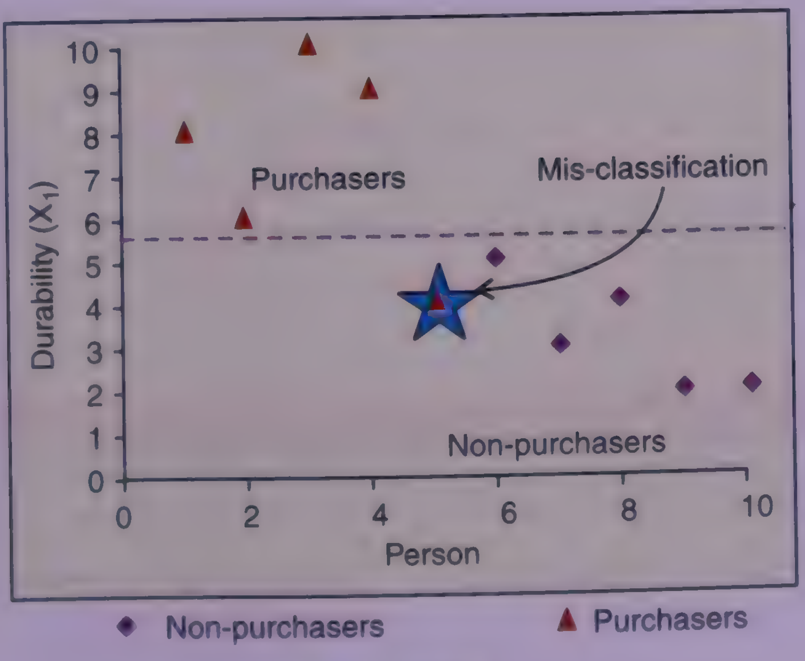


Fig. 16.1

Fix a cut-off score, say 5.5. Persons corresponding to the value of X_1 below 5.5 belong to no purchasers. On the other hand, persons corresponding to the value of X_1 above 5.5 belong to purchase Thus while, using only Durability (X_1) for discriminating between purchasers and non-purchase we have one misclassified person, person no. 5.

Similarly, when we use only Performance (X_2) for discrimination between two groups, we g two misclassifications, person no. 4 and 7 (we use cut-off score as 5.5, there is no better cut-o score giving less number of misclassifications); see Fig. 16.2.

	Purchasers					*Non-purchasers*				
Person	1	2	3	4	5	6	7	8	9	10
X_2	9	7	6	*4*	6	4	*7*	5	4	2

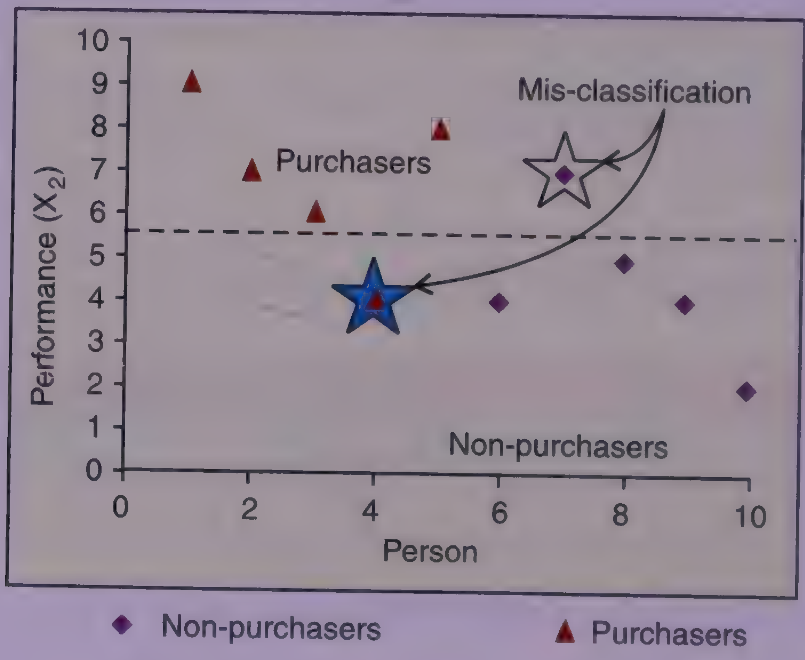


Fig. 16.2

When we use style (X_3) for discrimination purpose, we do not get a clear idea of discriminatin between two groups; see Fig. 16.3.

	Purchasers					*Non-purchasers*				
Person	1	2	3	4	5	6	7	8	9	10
X_3	6	5	3	4	2	7	2	5	3	2

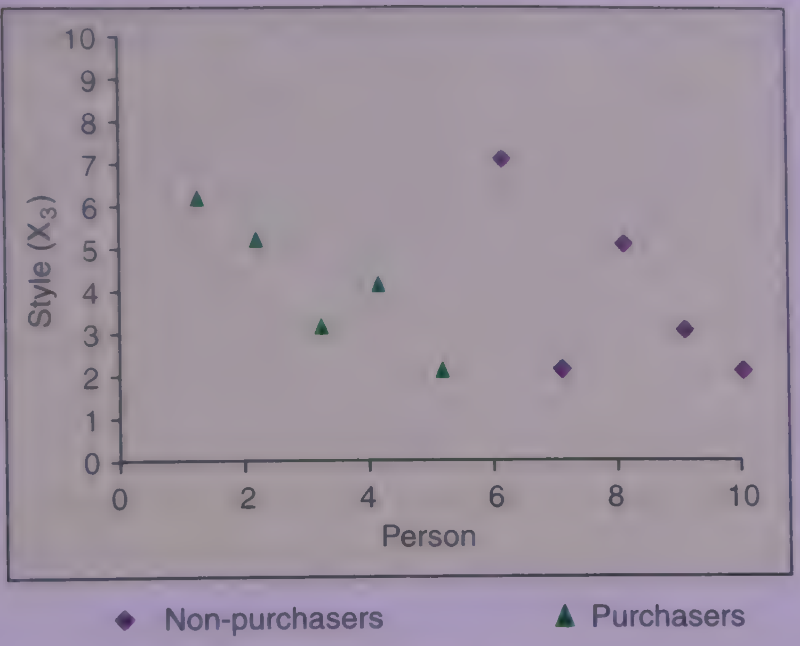


Fig. 16.3

Thus, X_1 is the best discriminating variable. While X_3 is the poorest in discriminating.

6.2 TWO GROUP DISCRIMINANT ANALYSIS

here are two popular methods of discriminant analysis — Fisher's method and Mahalanobis' method. oth of the methods are equivalent for two group discriminant analysis. We here present only the sher's method which is based on the idea of discriminant score obtained using a linear combination explanatory variables.

Let us revisit Example 16.1, where we have seen that X_3 is the poorest in discriminating variable. e leave X_3 and use X_1 and X_2 for discrimination between the two groups.

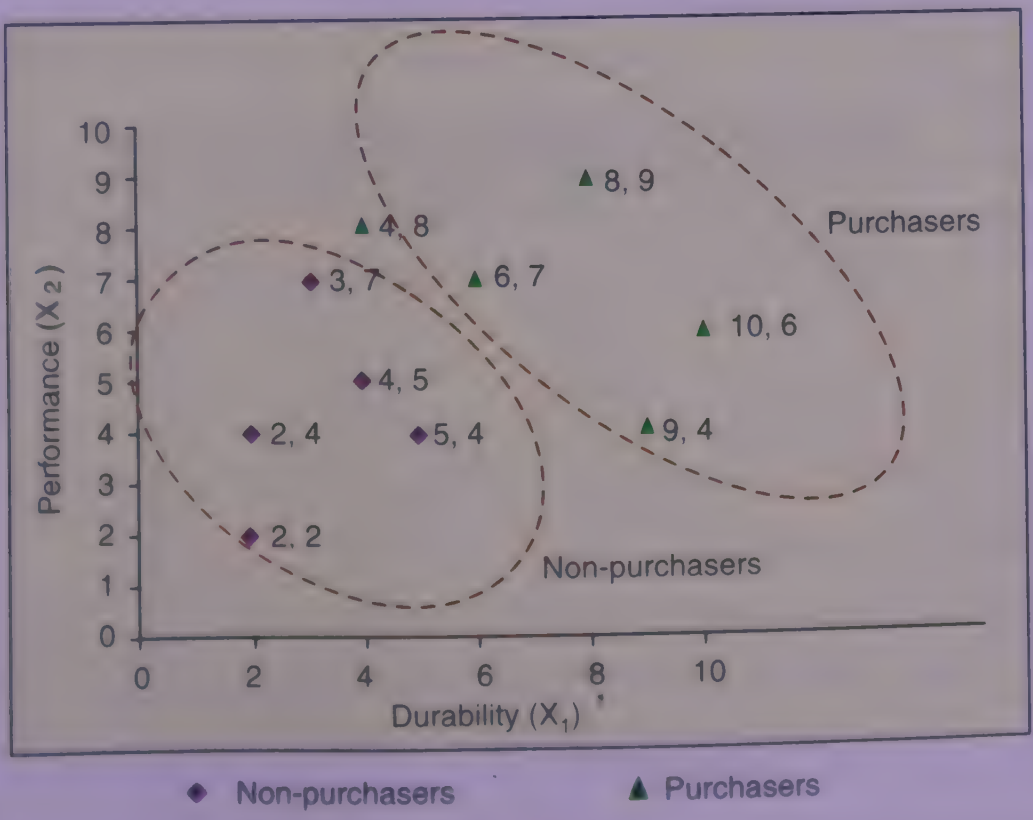


Fig. 16.4

Figure 16.4 suggests us to use a linear combination of X_1 and X_2 for discrimination. For examp if we take $(X_1 + X_2)$ for discrimination with a cut-off score 11, i.e., the individuals, having $(X_1 +$ <11, belong to the group of non-purchasers and the others belong to the group of purchasers. Tilt the axes graphically to separate the two groups may also help. How to obtain a linear combination explanatory variables which discriminates best between the groups is the important question wh we need to answer now.

16.2.1 Methodology

We wish to have a discriminant function, a linear combination of discriminating or explanatory variabl which maximize the difference between the two groups. We fit a discriminant function of the fo

$$Z = b_1 X_1 + b_2 X_2 + \dots + b_k X_k,$$

where Z is discriminant score, $X_1, X_2, \dots, X_k$ are k explanatory or predictor variables and b_j discriminant weight corresponding to variable X_j ($j=1, 2, \dots, k$).

We collect n observations on each variable. So, we can write

$$Z_i = b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + \dots + b_k x_{ki}$$

where, Z_i = Discriminant score ($i=1, 2, \dots, n$), and

x_{ij} = ith observation or value belonging predictor or explanatory variable X_i.

In a particular group, each individual has a discriminant score (Z_i). Discriminant weights b_j a obtained such that the discriminant function discriminates maximum between two groups.

Let **b** be the column vector of discriminant weights b_j and let $\mathbf{d} = (\bar{x}_{(2)} \times \bar{x}_{(1)})$ be the colum vector of the difference between the two group means (mean vectors of groups 1 and 2 are denot by $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ respectively). Also, let **C** be the pooled within-group covariance matrix given by **C** $(\mathbf{W}_1 + \mathbf{W}_2)/(n-2)$, where $\mathbf{W}_i$ is the matrix formed by sum of square & cross product terms of me corrected values of predictor variable belonging to group i ($i = 1, 2$).

We obtain weights b_j so as to maximize the ratio of between group variability to within grou variability. We obtain **b** such that

$$\frac{b' dd' b}{b' Cb}$$

is maximized. The numerator $b' dd' b$ captures the between group variability, while the denominat $b'Cb$ captures within group variability.

Centroid of a group is the mean of Z_i's of all objects within that group. A two group discrimina analysis has two centroids. It indicates most typical location of an individual from a particular grou Centroids of group 1 and group 2 are given by $\bar{Z}_{(1)} = \bar{x}_{(1)}'b$ and $\bar{Z}_{(2)} = \bar{x}_{(2)}'b$, respectively. Th greater the difference between the two centroids, larger the objective function $b'dd'b/b'Cb$ valu

It can be shown that such obtained b is proportional to $C^{-1}d$. Because the scale of discriminant ights is not known, we must standardize the discriminant weights such that the length of vector b

1. For this, we divide each b_j by $\sqrt{\sum_{j=1}^{k} b_j}$.

After obtaining discriminant weights, we can obtain the discriminant scores for n individuals in e sample. Suppose, n_1 = No. of persons or individuals in first group, and n_2 = No. of persons or dividuals in first group. Clearly, $n_1+n_2=n$. The cut-off score is obtained as $_1\bar{Z}_{(1)} + n_2\bar{Z}_{(2)})/(n_1 + n_2)$. When the discriminant score of an individual is below the cut-off score, e individual is assigned to group 1, otherwise to group 2.

.2.2 Assumptions

esults of discriminant analysis are valid only if

(i) Cases or individuals should be independent.

(ii) Predictor variables should have a multivariate normal distribution.

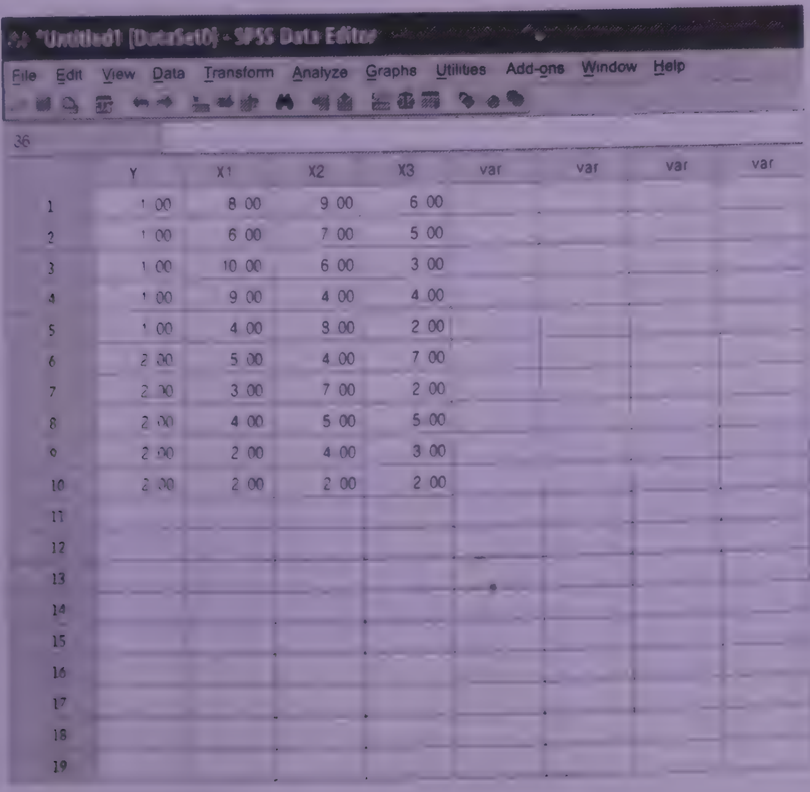
(iii) Within-group variance-covariance matrices should be equal across groups.

(iv) Group membership is assumed to be mutually exclusive, i.e., no case belongs to more than one group.

(v) Group memberships should be collectively exhaustive, i.e., all cases are members of a group.

6.2.3 Using SPSS

e use Example 16.1 to illustrate the usage of SPSS for discriminant analysis. We first enter the ta on three variables along with another variable for group membership. This new variable, denoted Y, takes the value 1 for group1, and the value 2 for group 2.



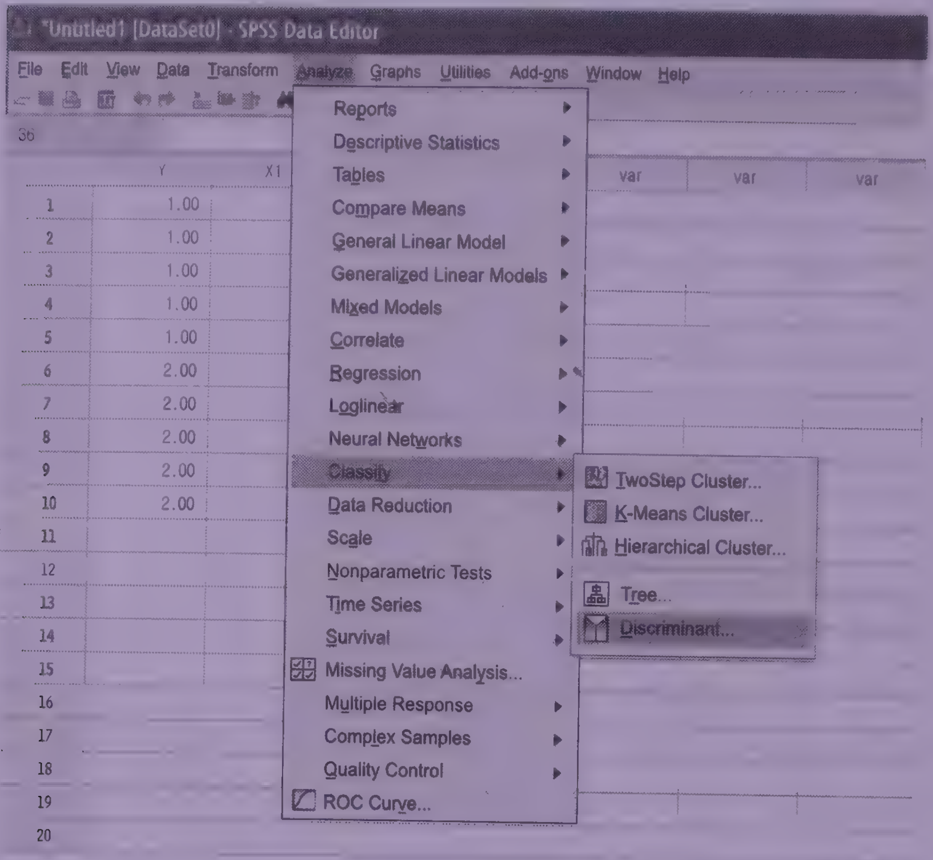
*Untitled1 [DataSet0] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

36

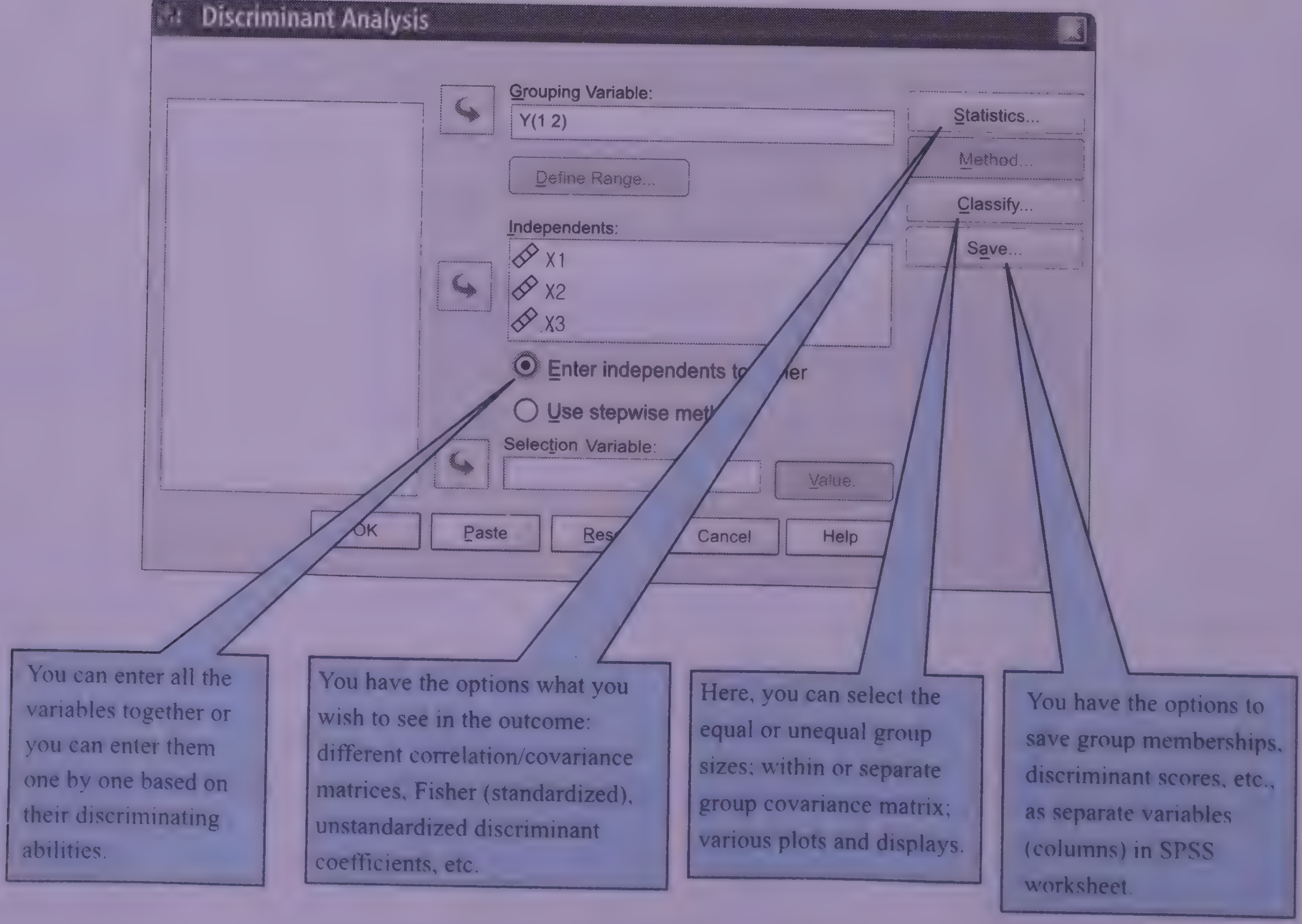
	Y	X1	X2	X3	var	var	var	var
1	1.00	8.00	9.00	6.00				
2	1.00	6.00	7.00	5.00				
3	1.00	10.00	6.00	3.00				
4	1.00	9.00	4.00	4.00				
5	1.00	4.00	3.00	2.00				
6	2.00	5.00	4.00	7.00				
7	2.00	3.00	7.00	2.00				
8	2.00	4.00	5.00	5.00				
9	2.00	2.00	4.00	3.00				
10	2.00	2.00	2.00	2.00				
11								
12								
13								
14								
15								
16								
17								
18								
19								

We click on 'Analyze', then select 'Classify' from the menu and then select 'Discriminant'.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...
Y X1 var var var

	Y
1	1.00
2	1.00
3	1.00
4	1.00
5	1.00
6	2.00
7	2.00
8	2.00
9	2.00
10	2.00

In the pop-up window, select *Y* as the 'Grouping variable'. Define the range: Minimum = Maximum = 2, for 2 groups. Select X$_1$ and X$_2$ as 'Independents'.


Discriminant Analysis
Define Range...
Statistics...
Method...
Classify...
Save...
Value...
OK
Paste
Reset
Cancel
Help

Grouping Variable: Y(1 2)

Independents: X1 X2 X3

● Enter independents together

○ Use stepwise method

Selection Variable:

You can enter all the variables together or you can enter them one by one based on their discriminating abilities.

You have the options what you wish to see in the outcome: different correlation/covariance matrices, Fisher (standardized), unstandardized discriminant coefficients, etc.

Here, you can select the equal or unequal group sizes; within or separate group covariance matrix; various plots and displays.

You have the options to save group memberships, discriminant scores, etc., as separate variables (columns) in SPSS worksheet.

The output is presented and explained as below:

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.315^a	100.0	100.0	0.877

[a] First 1 canonical discriminant functions were used in the analysis

Eigenvalue: These eigenvalues describe how much discriminating ability a function possesses. e magnitudes of the eigenvalues are indicative of the function's discriminating abilities. In case of o group discriminant analysis, you have only one eigenvalue.

% of Variance: This is the proportion of discriminating ability of predicting variables found in a ven function. This proportion is calculated as the proportion of the function's eigenvalue to the sum all the eigenvalues. In case of two group discriminant analysis, you have this value as 100%.

Cumulative %: This is the cumulative proportion of discriminating ability. In case of two group scriminant analysis, this value is also 100%.

Canonical Correlation: These are the canonical correlations of our predictor variables and the oupings variable. Canonical correlation is discussed in Chapter 18.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	d.f.	Sig.
1	0.232	9.504	3	0.023

Test of Function(s): These are the functions included in a given test with the null hypothesis at the canonical correlations associated with the functions are all equal to zero. In two group scriminant analysis, we have only one function.

Wilks' Lambda: In two group discriminant analysis, it is given by (1-canonical correlation^2). In ulti-group discriminant analysis, it is given by the product of all such values. Large values of Wilk's mbda indicate that group means are not different, while the small values indicate that group means e different.

Chi-square: This is the Chi-square statistic calculated using Wilk' Lambda. The null hypothesis that the function has no discriminating ability.

d.f.: This is the degrees of freedom for Chi-square test.

Sig.: This is the p-value associated with the Chi-square test. Here, the p-value is 0.023, which is naller than 0.05. Therefore, we reject the null hypothesis at 5% level of significance and conclude at the function has significant discriminating ability.

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X_1	1.110
X_2	0.709
X_3	–0.564

Standardized Canonical Discriminant Function Coefficients: These are standardiz discriminant function coefficients or discriminant weights, b_1, b_2, and b_3.

Thus, the discriminant function can be written as

$$Z_i = 1.110\, x_{1i} + 0.709\, x_{2i} - 0.564\, x_{3i}$$

Structure Matrix

	Function
	1
X_1	0.666
X_2	0.394
X_3	0.032

Pooled within-groups correlations between discriminating variables and standardized canonic discriminant functions. Variables ordered by absolute size of correlation within function.

Structure Matrix: This gives the correlation of predictor variables with the discriminant functio Clearly, X_1 is the best discriminating variable having highest correlation with discriminant functio

Canonical Discriminant Function Coefficients

	Function
	1
X_1	0.573
X_2	0.379
X_3	–0.297
(Constant)	– 4.002

Unstandardized coefficients

Canonical Discriminant Function Coefficients: As mentioned, you get here unstandardize discriminant weights with a constant term. SPSS obtains discriminant scores using unstandardize coefficients as

$$Z_i = 0.573\, x_{1i} + 0.379\, x_{2i} - 0.297\, x_{3i} - 4.002.$$

Adding a constant term (–4.002) makes the cut-off score as zero.

Functions at Group Centroids

	Function
Y	1
1	1.629
2	–1.629

Unstandardized canonical discriminant functions evaluated at group means.

Functions at Group Centroids: Group centroids $\bar{Z}_{(1)} = 1.629$ and $\bar{Z}_{(2)} = -1.629$ using nstandardized coefficients. Also, the cut-off score, using unstandardized coefficients, is zero.

One may prefer to use standardized coefficients because the scale of the discriminant function s not known.

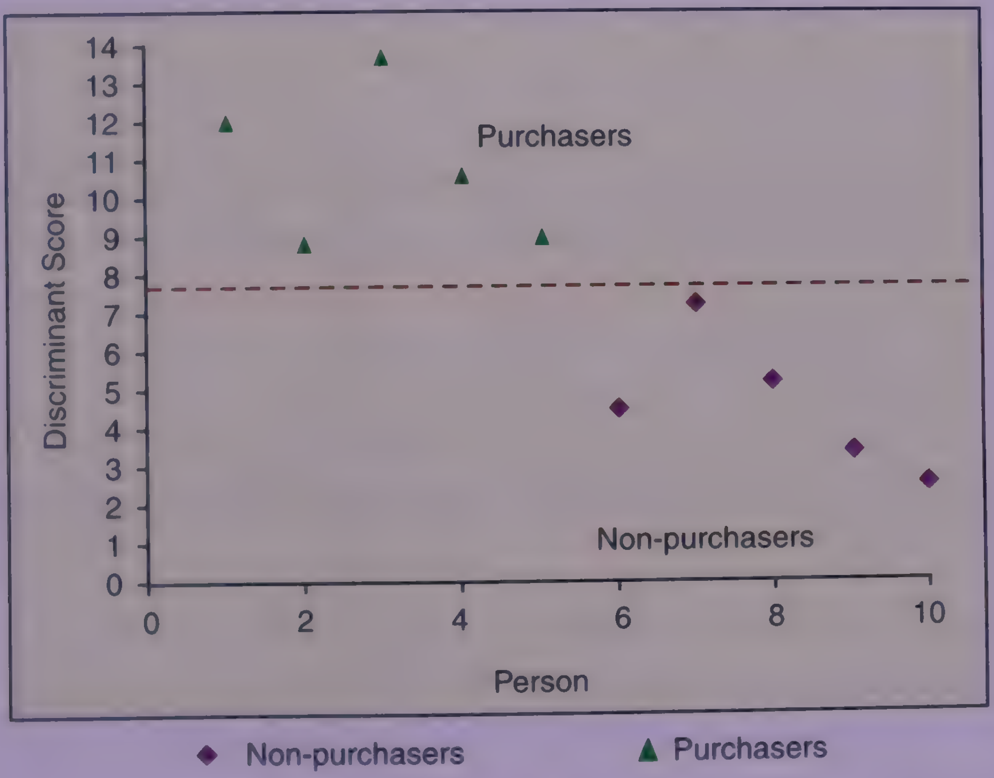


Fig. 16.5

Note: Sometimes, classification matrix is also constructed to show the group membership. In case of no misclassification, all diagonals are non-zero and all off-diagonals are zero.

16.2.4 Predicting Group Membership

The discriminant function, we fit in the previous section is

$$Z_i = 1.110\, x_{1i} + 0.709\, x_{2i} - 0.564\, x_{3i}$$

Using this discriminant function, we obtain the discriminant scores as

11.877	8.803	13.662	10.57	8.984	4.438	7.165	5.165	3.364	2.51

First 5 scores belong to group 1 and last 5 scores belong to group 2. The group centroids a calculated as 10.7792 and 4.5284. The cut-off score (10.7792 + 4.5284)/2 = 7.6538.

Suppose an individual assigns the ranks 3, 5, and 9 to durability (X_1), performance(X_2), and sty (X_3), respectively. To which group he should be assigned purchasers or non-purchasers?

To answer this question, we obtain the discriminant score for this individual.

The discriminant score $Z = 1.799$, which is smaller than the cut-off score 7.6538. Therefore, th new individual is assigned to group 1, i.e., non-purchasers.

16.3 MULTIPLE DISCRIMINANT ANALYSIS

When we need to discriminate among more than two groups, we use multiple discriminant analysi This technique requires fitting g-1 number of discriminant functions, where g is the number of group

16.3.1 Methodology

For multiple discriminant analysis, we use the similar approach of Fisher. We maximize the ration among groups or across group variability and within group variability. We obtain b such that

$$\lambda = \frac{b'Ab}{b'Wb'}$$

is maximum. Here **A** is the across group sum of squares matrix and **W** is within group sum of square matrix. When we solve this maximization problem using differential calculus, it reduces to th problem of solving the equation

$$W^{-1}Ab = \lambda b$$

Clearly, λ gives the eigenvalues of the matrix $W^{-1}A$ and b becomes corresponding eigenvectors As earlier, we need to standardize the vector b so that the length of b becomes unity. The techniqu provides valid results under the same assumptions we made in two group discriminant analysis.

16.3.2 Using SPSS

We illustrate the usage of SPSS through following example:

Example 16.2: A manufacturer of snacks is considering launching a new banana chips in the market. He wishes to study the perception of consumers regarding three popular brands of banana chips 'A', 'B', 'C'. Data was collected from 15 canteen owners, taking five for each brand of banana chips, on five variables:

(i) Price (measured as a rank, 0–5, higher rank is favorable).

(ii) Quality (measured as rank, 0–25, higher rank is favorable).

(iii) Availability (number of packets purchased by canteen owner in last one week duration, 0–500, higher rank is favorable).

(iv) Packing (measured as rank, 0–5, higher rank is favorable).

(v) Taste (measured as rank, 0–20, higher rank is favorable).

The collected data is shown below:

Brand	Price (0–5)	Quality (0–25)	Availability (0–500)	Packaging (0–5)	Taste (0–20)
A	3	20	500	5	18
A	2	21	234	4	15
A	1	22	250	4	14
A	4	8	345	5	12
A	3	9	432	3	13
B	1	1	234	2	15
B	2	3	231	3	11
B	5	9	45	4	10
B	4	1	35	3	12
B	3	2	25	2	10
C	1	8	75	4	8
C	3	10	80	4	7
C	4	14	90	5	10
C	2	3	96	2	12
C	2	4	59	2	6

Help the manufacturer in studying consumer's perception and discriminate between the three rands.

olution: We use SPSS as in the previous example. Some important outcomes are discussed below:

Standardized canonical discriminant function coefficients

	Function	
	1	2
Price	0.317	0.610
Quality	1.062	–0.464
Availability	1.045	–0.188
Packaging	–0.492	–0.261
Taste	–0.170	1.019

In 3 group discriminant analysis, we fit two functions. Standardized discriminant coefficients weights are given in above table.

Function 1:

$Z = 0.317$ Price + 1.062 Quality + 1.045 Availability – 0.492 Packaging – 0.170 Taste

Function 2:

$Z = 0.610$ Price – 0.464 Quality – 0.188 Availability – 0.261Packaging +1.019 Taste

Structure Matrix

	Function	
	1	2
Availability	0.652*	0.284
Quality	0.496*	–0.338
Packaging	0.260*	–0.229
Taste	0.428	0.694*

Pooled within-groups correlations between discriminating variables and standardize canonical discriminant functions. Variables ordered by absolute size of correlation withi function.

*Largest absolute correlation between each variable and any discriminant function

Functions at Group Centroids

Group	Function	
	1	2
1	2.811	0.090
2	–1.584	1.019
3	–1.227	–1.109

Unstandardized canonical discriminant functions evaluated at group means.

The meanings of structure matrix and group centroids have already been explained.

Figure 16.6 plots the discriminant scores. It is clear that Function 1 discriminates well betwee Group 1 and Groups 2 and 3, while Function 2 can be used to discriminate between Group 2 an Group 3. However, the discriminating ability of Function 2 is not very good.

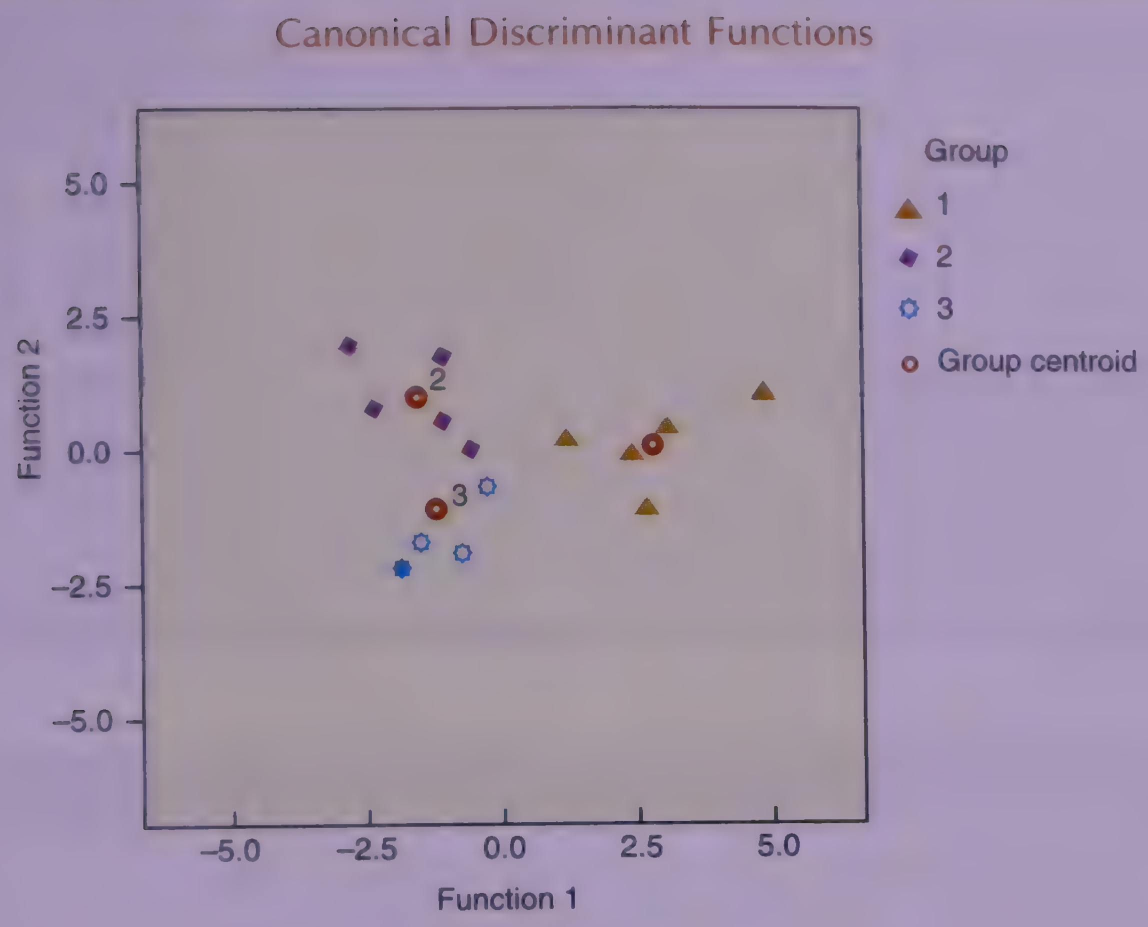


Fig. 16.6

Using the discriminant scores of the two functions, we can obtain the cutoff scores for Function and Function 2 as well. These cut-off scores are used to predict the group membership of a new dividual.

PROBLEMS

1. What is the meaning of discrimination between groups?
2. What do you understand by discriminating ability of a variable?
3. Compare regression analysis with discriminant analysis?
4. Explain the difference between two group discriminant analysis and multiple discriminant analysis.
5. How many discriminant functions would you obtain in a 4 group discriminant analysis and why?
6. Following is the partial outcome of a discriminant analysis problem solved using SPSS.

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*d.f.*	*Sig.*
1	---	----	---	0.23

What do you interpret from this outcome?

7. What is the importance of Wilk`s lambda? Canonical correlation in a two group discriminant analysis is 0.6. Obtain Wilk's lambda.
8. What do you mean by misclassification in discriminant analysis?

17

Chapter

Cluster Analysis

17.1 INTRODUCTION

The purpose of cluster analysis is to divide large group of objects or observations, like customers products, into smaller groups such that the observations within each group are similar or close (homogeneous) and the observations in different groups are dissimilar or far away. These small groups are called as clusters. Thus, resulting clusters exhibit high internal (within cluster) homogenei and high external (between clusters) heterogeneity.

Cluster analysis is an inter-dependence technique and makes no distinction between depende (study) and independent (explanatory) variables. Note that, we do not combine the variables here done in factor analysis. In factor analysis, we reduce number of variables by grouping them in smaller set of factors, while in cluster analysis we reduce number of observations by grouping the into smaller set of clusters. Number of groups (or clusters) is not known in advance; they are natur clusters as suggested by the data. Main applications of cluster analysis are segmenting the mark and determining target markets, product positioning and new product development, and selecting te markets. Some applications of cluster analysis are illustrated below:

(i) Identifying people with similar patterns of past purchases so that you can tailor your marketi strategies.

(ii) Clubbing television shows into homogeneous categories based on viewer characteristic (market segmentation)

(iii) Examining patients with a diagnosis of swine flu to determine if distinct subgroups can identified based on a symptom checklist and results from pathological tests.

In cluster analysis, we group the observations based on a particular set of variables. Thes variables affect the clusters obtained using cluster analysis. There is no general rule of selecting th variables. We have to be careful while selecting these variables. After selecting the variables, w obtain sample observations on those variables.

7.2 DISTANCE MEASURES

cluster analysis, we group the observations on the basis of proximity or similarity. There are eral different measures of distance or proximity. Some important measures are discussed below:

(i) **Euclidean Distance:** Suppose, there are k variables denoted as $X_1, X_2, \ldots, X_k$. The euclidean distance between two observations $(x_{i1}, x_{i2}, \ldots, x_{ik})$ and $(x_{j1}, x_{j2}, \ldots, x_{jk})$ is given by

$$d_{ij} = \left[\sum_{h=1}^{k} \left(x_{ih} - x_{jh}\right)^2\right]^{1/2}$$

(ii) **Manhattan Distance:** between two observations $(x_{i1}, x_{i2}, \ldots, x_{ik})$ and $(x_{j1}, x_{j2}, \ldots, x_{jk})$ is given by

$$\sum_{h=1}^{k} \left|x_{ih} - x_{jh}\right|$$

(iii) **Chebyshev Distance:** between two observations $(x_{i1}, x_{i2}, \ldots, x_{ik})$ and $(x_{j1}, x_{j2}, \ldots, x_{jk})$ is

$$MAX_h \left|x_{ih} - x_{jh}\right|$$

(iv) **Mahalanobis or correlation Distance:** between two observations $x_i' = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and $x_j' = (x_{j1}, x_{j2}, \ldots, x_{jk})$ is

$$D_{ij} = (x_i - x_j)' \sum^{-1} (x_i - x_j)$$

here, $\sum$ is population covariance matrix of the variables $(X_1, X_2, \ldots, X_k)$.

(v) **Correlation:** between two observations $(x_{i1}, x_{i2}, \ldots, x_{ik})$ and $(x_{j1}, x_{j2}, \ldots, x_{jk})$ is given by

$$r_{ij} = \frac{\sum_{h=1}^{n} (x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j)}{\sqrt{\left[\sum_{h=1}^{n} (x_{ih} - \bar{x}_i)^2\right]\left[\sum_{h=1}^{n} (x_{jh} - \bar{x}_j)^2\right]}}$$

here, $\bar{x}_i$ and $\bar{x}_j$ are corresponding means.

(vi) **Matching measure:** is used when the observations are categorical. It is given by number of matches divided by number of variables. For example, consider the following data:

Person	Smoker (Yes = 1, No = 0)	Gender (Female = 1, Male = 0)	Adult (Yes = 1, No = 0)	Graduate (Yes = 1, No = 0)
1	0	1	1	1
2	1	0	0	0
3	1	1	0	0
4	1	1	1	0
5	1	1	1	0

In this data, distance between person 1 and person 2 is given by number of matches divided number of variables, which is 0/4 or 0. We can write all such distances as below:

Person	1	2	3	4	5
1	1				
2	0/4	1			
3	1/4	3/4	1		
4	2/4	2/4	3/4	1	
5	2/4	2/4	3/4	4/4	1

Euclidean, Manhattan, and Chebyshev distances depend upon units of measurement. In ca units of measurement differ too much it is advisable to standardize each variable. This is done avoid giving too much weight to variables whose scale is larger. For standardization, we subtract th mean and divide by the standard deviation. For example, to standardize the observations on X_1, give by $(x_{11}, x_{12}, ..., x_{1n})$, we obtain mean $\overline{x}_1$ and standard deviation s_1 of these observations an transform each observation using $(\overline{x}_{1i} - \overline{x}_1)/s_1$ for i=1, 2,..., n. These standardized observation are called as Z-scores.

17.3 CLUSTERING ALGORITHMS

There are several clustering algorithms. The popular algorithms are divided into two types Non hierarchical clustering and hierarchical clustering. These methods are explained below.

17.3.1 Non-hierarchical Clustering

This method is also known as *k*-mean clustering. The steps involved in this method are describe below:

(a) Fix *k*, number of clusters in advance.

(b) Take the initial guess of *k* cluster centroids or means. The popular approach of this is t take the observations (or cases) having highest distances.

(c) Proceed through the list of cases assigning a case to the cluster whose centroid is th nearest. Recalculate the centroid is the nearest. Recalculate the centroid of new cluste receiving the new case and the cluster losing an old case.

(d) Repeat step (c) until the reassignment of cases occurs.

The final assignment of cases to clusters depends on the initial guess of *k* means. In order t check the stability of the clustering, it is desirable to run the algorithm again with a new initial guess

How to select the number of clusters *k* is a big question. It is decided by the researcher wha value of *k* yields the "best" solution. We suggest repeating the procedure with different values of and comparing the solutions. When the number of clusters is smaller, the solution looks simpler However the adequacy of the solution may increase with the number of clusters.

k-mean clustering is affected highly by outliers, since they will usually be selected as initial luster centers. This will result in outliers forming clusters with small numbers of cases. Therefore, e suggest the researcher to screen the data for outliers before starting a cluster analysis.

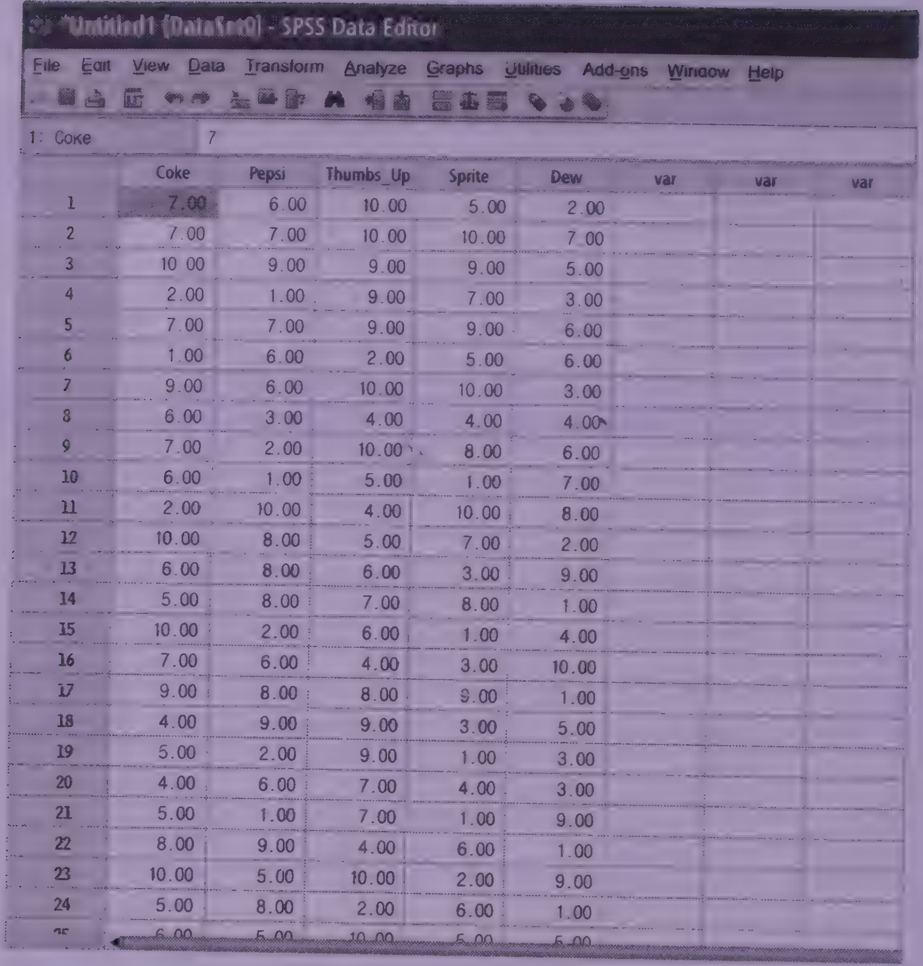
xample 17.1: Following data shows the preference of 25 students of 5 varieties of carbonated oft drinks (softdrink.xls):

Student	*Coke*	*Pepsi*	*Thumbs Up*	*Sprite*	*Dew*
1	7	6	10	5	2
2	7	7	10	10	7
3	10	9	9	9	5
4	2	1	9	7	3
5	7	7	9	9	6
6	1	6	2	5	6
7	9	6	10	10	3
8	6	3	4	4	4
9	7	2	10	8	6
10	6	1	5	1	7
11	2	10	4	10	8
12	10	8	5	7	2
13	6	8	6	3	9
14	5	8	7	8	1
15	10	2	6	1	4
16	7	6	4	3	10
17	9	8	8	9	1
18	4	9	9	3	5
19	5	2	9	1	3
20	4	6	7	4	3
21	5	1	7	1	9
22	8	9	4	6	1
23	10	5	10	2	9
24	5	8	2	6	1
25	6	5	10	5	5

Conduct *k*-mean clustering with some suitable value of *k* and obtain homogeneous clusters. Help us understand the differences between clusters.

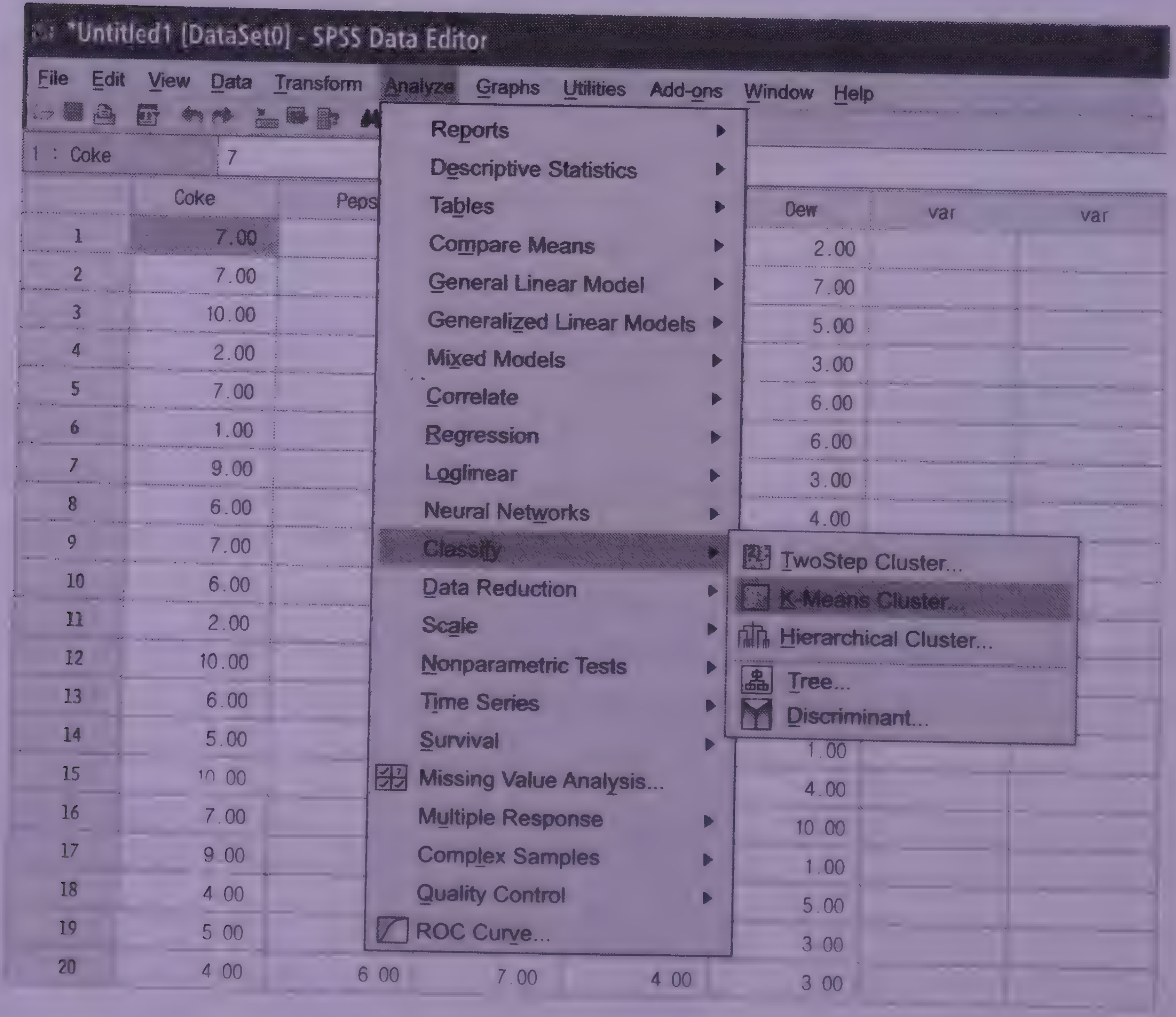
Solution: We use SPSS as below:

We first enter the data in SPSS worksheet.

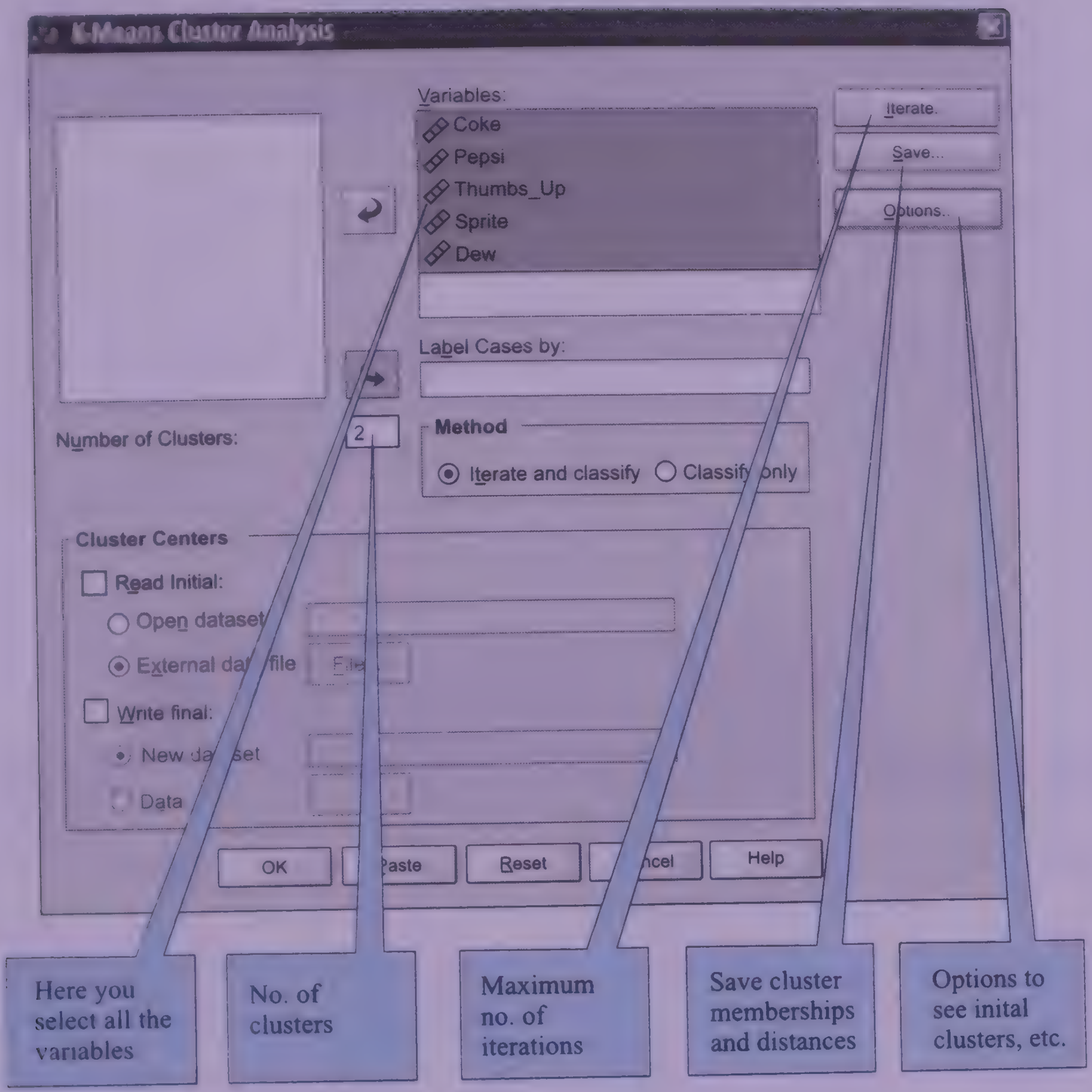

*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : Coke 7

	Coke	Pepsi	Thumbs_Up	Sprite	Dew	var	var	var
1	7.00	6.00	10.00	5.00	2.00			
2	7.00	7.00	10.00	10.00	7.00			
3	10.00	9.00	9.00	9.00	5.00			
4	2.00	1.00	9.00	7.00	3.00			
5	7.00	7.00	9.00	9.00	6.00			
6	1.00	6.00	2.00	5.00	6.00			
7	9.00	6.00	10.00	10.00	3.00			
8	6.00	3.00	4.00	4.00	4.00			
9	7.00	2.00	10.00	8.00	6.00			
10	6.00	1.00	5.00	1.00	7.00			
11	2.00	10.00	4.00	10.00	8.00			
12	10.00	8.00	5.00	7.00	2.00			
13	6.00	8.00	6.00	3.00	9.00			
14	5.00	8.00	7.00	8.00	1.00			
15	10.00	2.00	6.00	1.00	4.00			
16	7.00	6.00	4.00	3.00	10.00			
17	9.00	8.00	8.00	9.00	1.00			
18	4.00	9.00	9.00	3.00	5.00			
19	5.00	2.00	9.00	1.00	3.00			
20	4.00	6.00	7.00	4.00	3.00			
21	5.00	1.00	7.00	1.00	9.00			
22	8.00	9.00	4.00	6.00	1.00			
23	10.00	5.00	10.00	2.00	9.00			
24	5.00	8.00	2.00	6.00	1.00			

Click 'Analyze', select 'Classify', select 'K-means Cluster'.


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : Coke 7
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...

A window pops up.


K-Means Cluster Analysis
Iterate...
Save...
Options...
Label Cases by:
Cluster Centers
☐ Read Initial:
○ Open dataset
● External data file
File...
☐ Write final:
● New dataset
○ Data file
OK
Paste
Reset
Cancel
Help
Here you select all the variables
No. of clusters
Maximum no. of iterations
Save cluster memberships and distances
Options to see inital clusters, etc.

Variables:
- Coke
- Pepsi
- Thumbs_Up
- Sprite
- Dew

Number of Clusters: 2

Method: ● Iterate and classify ○ Classify only

We get the following outcome:

Initial cluster centers

	Cluster	
	1	2
Coke	5.00	8.00
Pepsi	1.00	9.00
Thumbs Up	7.00	4.00
Sprite	1.00	6.00
Dew	9.00	1.00

Iteration History[a]

Iteration	Change in cluster centers	
	1	2
1	4.439	4.626
2	0.000	0.000

[a] Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 13.077.

Final cluster centers

	Cluster	
	1	2
Coke	6.36	6.29
Pepsi	3.27	7.64
Thumbs Up	7.27	6.86
Sprite	3.27	7.21
Dew	6.27	3.64

Number of coses in each cluster

	Cluster	
	1	2
Valid	11.000	14.000
Missing	25.000	0.000

Outcomes are easy to interpret. From 'Final Cluster Centers', it is clear that the preference of Coke and Thumbs Up are almost the same in both the clusters.

7.3.2 Hierarchical Clustering

here are two methods of hierarchical clustering — divisive and agglomerative.

In divisive method. all the cases or observations are first put in one cluster. At each stage the luster is divided into more clusters. It ends up having as many clusters as the cases are. We start aving one cluster and end with n clusters, where n is the number of cases or observations. In this ethod, once a cluster is divided into two or more clusters, they can not be merged.

In agglomerative method, we first consider each case as a cluster. Then closest (most similar) lusters are combined (agglomerated). It ends up having only one cluster of all the cases. Thus, we tart with n clusters and end with one cluster. Once the clusters are merged, they cannot be split.

In these methods, first and last iterations are not the best solutions. We need to know the number f clusters we require. There is no right or wrong answer as to how many clusters we should have. o find a good number of clusters, we should study the solution at each step and decide a reasonable umber of homogeneous clusters representing the data.

We also require a suitable distance measure and a criterion to decide which clusters are divided r merged at successive steps.

17.4 AGGLOMERATIVE CLUSTERING

n this section, we describe agglomerative clustering procedure in detail. For agglomerative clustering, e need

(i) A suitable distance measure
(ii) A criterion to combine the clusters
(iii) Number of clusters

We have already discussed about some distance measures in Section 17.2. It is advisable to tandardize the data so that no variable is under/over-represented. Below, we discuss about various riteria of combining the clusters.

17.4.1 Combining Clusters

n Agglomerative hierarchical clustering, we start with each case being a cluster. At the next step, he two cases having the smallest distance are joined into a single cluster. At every step, we combine

(i) two individual cases, or
(ii) an individual case to an existing cluster, or
(iii) two existing clusters.

The distance between individual cases is unambiguous; it is the same as the value of distance easure. However. when we need to combine two clusters or a cluster and a case, we need to efine the distance between two clusters. Some popular criteria of combining clusters or measures f distance between two clusters are given below:

(i) **Nearest neighbor (single linkage):** It is defined as the smallest distance between two cases in the different clusters.

(ii) **Furthest neighbor (complete linkage):** The distance between two clusters is define as the distance between the two furthest points.

(iii) **UPGMA (unweighted pair-group method using arithmetic averages):** This is define as the average of the distances between all pairs of cases in which one member of the pai is from each of the clusters. Since, it uses information about all pairs of distances, it i preferred to the single linkage and complete linkage.

(iv) **Ward's method:** First, we calculate the mean of all variables for each cluster. Then, fo each case, the squared Euclidean distance to the cluster means is calculated. These distance are summed for all of the cases. At each step, those two clusters are combined which result in the smallest increase in the overall sum of the squared within-cluster distances.

(v) **Centroid method:** First, we calculate the mean of all variables for each cluster. The distance between two clusters is given by the sum of distances between cluster means.

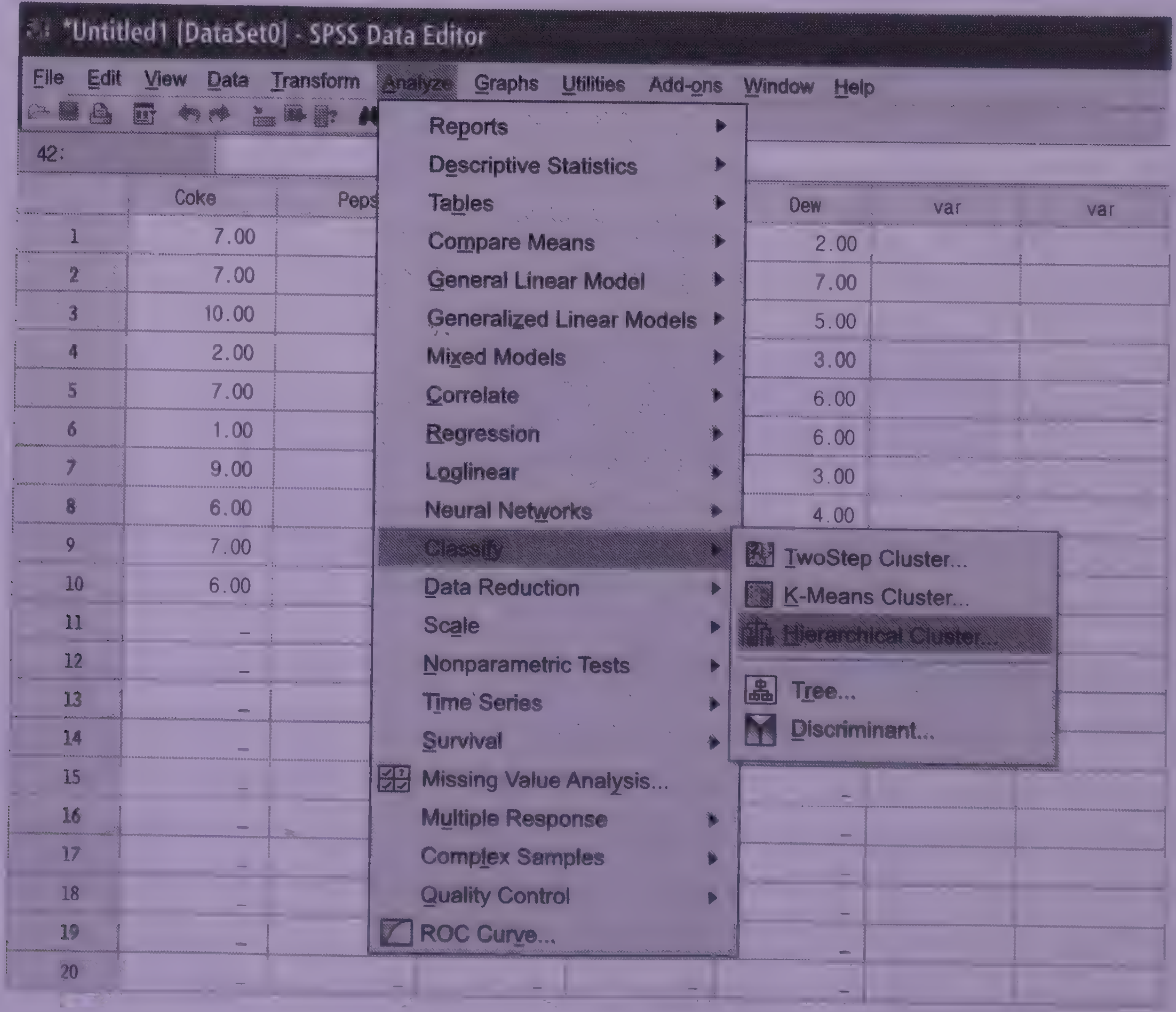
17.4.2 Using SPSS

We illustrate the usage of SPSS and explain the outcomes through the following example:

Example 17.2: Consider the data given in Example 17.1 and obtain homogeneous clusters using agglomerative clustering.

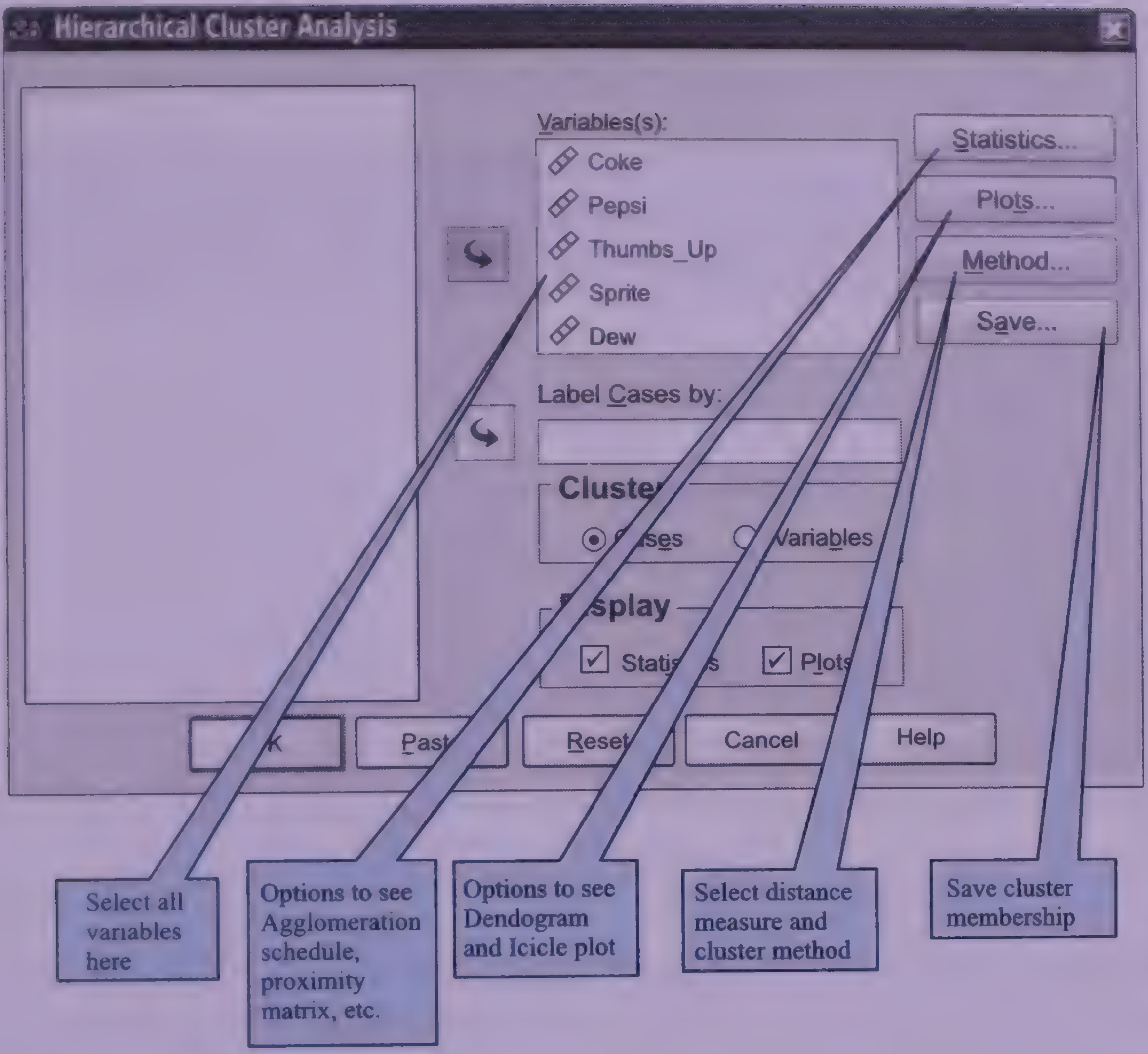
Solution: It is difficult to show the outcome consisting 25 cases here, so we consider only first 10 cases in the given data.

We select 'Hierarchical Cluster' as shown below:


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
42:
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...

	Coke	Peps	Dew	var	var
1	7.00		2.00		
2	7.00		7.00		
3	10.00		5.00		
4	2.00		3.00		
5	7.00		6.00		
6	1.00		6.00		
7	9.00		3.00		
8	6.00		4.00		
9	7.00				
10	6.00				

Following window pops up:



Variables(s):
- Coke
- Pepsi
- Thumbs_Up
- Sprite
- Dew

Cluster: ● Cases ○ Variables

Display: ☑ Statistics ☑ Plots

Select all variables here

Options to see Agglomeration schedule, proximity matrix, etc.

Options to see Dendogram and Icicle plot

Select distance measure and cluster method

Save cluster membership

The outcome is given and explained as below:

Proximity Matrix

	Euclidean distance									
Case	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:Case 7	8:Case 8	9:Case 9	10:Case 10
1:Case 1	.000	3.277	2.650	2.693	2.652	4.108	1.912	2.586	2.830	3.961
2:Case 2	3.277	.000	1.772	3.724	.733	3.893	2.372	3.617	1.976	4.079
3:Case 3	2.650	1.772	.000	4.245	1.406	4.389	1.653	3.544	2.804	4.517
4:Case 4	2.693	3.724	4.245	.000	3.319	3.474	3.257	2.594	2.527	3.596
5:Case 5	2.652	.733	1.406	3.319	.000	3.483	1.916	3.007	1.835	3.743
6:Case 6	4.108	3.893	4.389	3.474	3.483	.000	4.598	2.485	3.863	3.084
7:Case 7	1.912	2.372	1.653	3.257	1.916	4.598	.000	3.278	2.405	4.612
8:Case 8	2.586	3.617	3.544	2.594	3.007	2.485	3.278	.000	2.717	2.106
9:Case 9	2.830	1.976	2.804	2.527	1.835	3.863	2.405	2.717	.000	2.989
10:Case 10	3.961	4.079	4.517	3.596	3.743	3.084	4.612	2.106	2.989	0.000

This is a dissimilarity matrix.

Proximity Matrix: The proximity matrix gives the distances between each pair of cases. Not that the diagonal elements are zero as a case has zero distance from itself.

Agglomeration Schedule

	Cluster combined		Coefficients	Stage cluster first appears		
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	Next Stage
1	2	5	.367	0	0	4
2	3	7	1.193	0	0	5
3	8	10	2.246	0	0	6
4	2	9	3.395	1	0	8
5	1	3	4.640	0	2	8
6	6	8	6.145	0	3	7
7	4	6	7.921	0	6	9
8	1	2	9.700	5	4	9
9	1	4	13.628	8	7	0

Agglomeration Schedule: In agglomeration schedule, we get the schedule of cluster agglomeration. In first step cluster 2 (case 2) and cluster 5 (case 5) are merged. The new cluster made by combining these two cases will be called as cluster 2 (smallest number). This cluster 2 is next time used in stage 4. In coefficients, we get the distance or measure obtained using cluster combining method (here we use Ward's method). These coefficients are used to decide how many clusters are needed to represent the data. We want to stop cluster formation when the increase in the Coefficients column between two adjacent steps is large. In the given example, we can stop cluster formation after 8th stage. Thus, we have 2 clusters.

In this example, we may like to stop at the four-cluster solution, after stage 3. We can write the schedule as below also:

Stage 1: (2) + (5) = (2, 5)
Stage 2: (3) + (7) = (3, 7)
Stage 3: (8) + (10) = (8, 10)
Stage 4: (2, 5) + (9) = (2, 5, 9)
Stage 5: (1) + (3, 7) = (1, 3, 7)
Stage 6: (6) + (8, 10) = (6, 8, 10)
Stage 7: (4) + (6, 8, 10) = (4, 6, 8, 10)
Stage 8: (1, 3, 7) + (2, 5, 9) = (1, 2, 3, 5, 7, 9)
Stage 9: (1, 2, 3, 5, 7, 9) + (4, 6, 8, 10) = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Here the bracket '( )' shows the cluster and the numbers inside the bracket shows cases in the cluster.

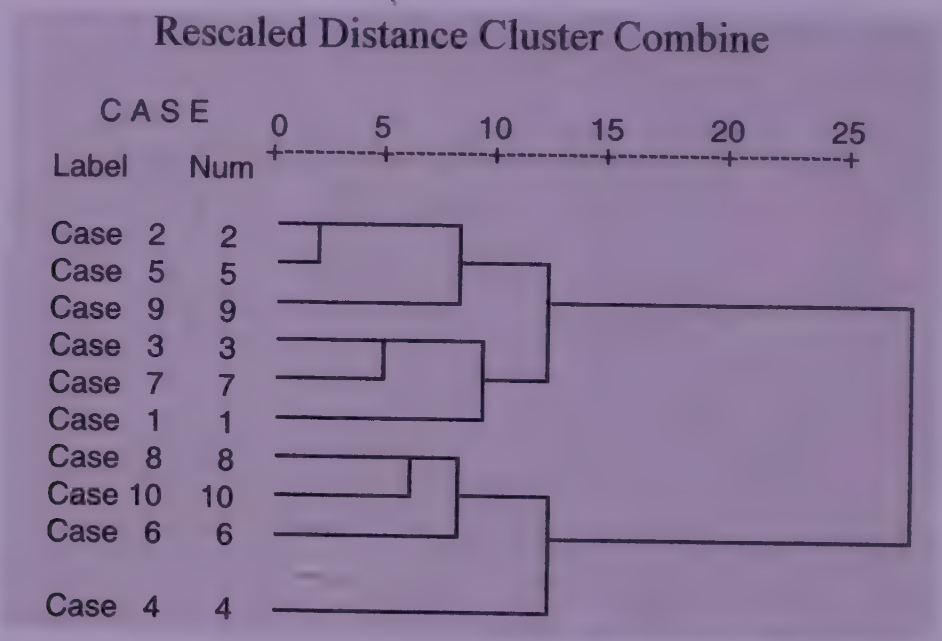
Vertical Icicle

Number of clusters	Case																		
	10: Case 10		8: Case 8		6: Case 6		4: Case 4		9: Case 9		5: Case 5		2: Case 2		7: Case 7		3: Case 3		1: Case 1
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X
4	X	X	X	X	X		X		X	X	X	X	X		X	X	X	X	X
5	X	X	X		X		X		X	X	X	X	X		X	X	X	X	X
6	X	X	X		X		X		X	X	X	X	X		X	X	X		X
7	X	X	X		X		X		X		X	X	X		X	X	X		X
8	X		X		X		X		X		X	X	X		X	X	X		X
9	X		X		X		X		X		X	X	X		X		X		X

Icicle Plot: Above plot is called as icicle plot which shows the schedule of agglomeration starts from the bottom. We have a column for each case having 'X' sign in each of its rows. We a have a column between two case-columns. Note the last row. When cluster 2 and cluster 5 a combined, we get a 'X' sign between these two case-columns in the last row.

Hierarchical Cluster Analysis

Dendrogram using Ward Method



Dendrogram: The word dendrogram is made by two Greek words 'dendron' (tree) and 'gramma (drawing). Dendrogram is used to illustrate the arrangement of the clusters produced by hierarchic clustering. It gives a visual representation of the distance at which clusters are combined. It is rea from left to right. Vertical lines show joined clusters. The positions of these lines show the distance which clusters are joined. These distances in the plot are not the actual distances; they are proportiona to the actual distances. The first vertical line corresponds to case 2 and case 5.

Based on this, we can decide having 2 clusters – cluster 1 having cases 2, 5, 9, 3, 7, and 1 an cluster 2 having cases 8, 10, 6, and 4.

PROBLEMS

1. What is the difference between cluster analysis and factor analysis?
2. What is the difference between discriminant analysis and cluster analysis?
3. When the cases or observations are measured on categorical scale, each variable having two categories, how will you define a distance measure?
4. Explain the difference between hierarchical clustering and non-hierarchical clustering.
5. What is agglomerative clustering? How it is different from divisive clustering?
6. What do you understand by Ward's method of combining clusters?
7. Write a short note on the followings:
   - (a) Dendrogram
   - (b) Icicle plot
   - (c) Agglomeration schedule
   - (d) Proximity Matrix
   - (e) Single linkage
   - (f) Complete Linkage

18

Chapter

ther Multivariate Techniques

e have already discussed some important multivariate techniques, viz., multiple regression analysis hapter 14), factor analysis (Chapter 15), cluster analysis (Chapter 16), and multiple discriminant alysis (Chapter 17).

All statistical techniques which simultaneously analyse more than two variables on a sample of servations can be categorized as multivariate techniques. We may as well use the term 'multivariate alysis' which is a collection of methods for analyzing data in which a number of observations are ailable for each object. In the analysis of many problems, it is helpful to have a number of scores each object. For instance, in the field of intelligence testing if we start with the theory that general elligence is reflected in a variety of specific performance measures, then to study intelligence in e context of this theory one must administer many tests of mental skills, such as vocabulary, speed recall, mental arithmetic, verbal analogies and so on. The score on each test is one variable, X_i, and ere are several, k, of such scores for each object, represented as X_1, X_2, …,X_k. Most of the search studies involve more than two variables in which situation analysis is desired of the association tween one (at times many) criterion variable and several independent variables, or we may be quired to study the association between variables having no dependency relationships. All such alyses are termed as multivariate analyses or multivariate techniques. In brief, techniques that e account of the various relationships among variables are termed multivariate analyses or ıltivariate techniques.

Of late, multivariate techniques have emerged as a powerful tool to analyse data represented in rms of many variables. The main reason being that a series of univariate analysis carried out parately for each variable may, at times, lead to incorrect interpretation of the result. This is so cause univariate analysis does not consider the correlation or inter-dependence among the variables. a result, during the last fifty years, a number of statisticians have contributed to the development several multivariate techniques. Today, these techniques are being applied in many fields such as onomics, sociology, psychology, agriculture, anthropology, biology and medicine. These techniques e used in analyzing social, psychological, medical and economic data, specially when the variables ncerning research studies of these fields are supposed to be correlated with each other and when gorous probabilistic models cannot be appropriately used. Applications of multivariate techniques in actice have been accelerated in modern times because of the advent of high speed electronic mputers.

In this chapter, we give a brief idea about some other multivariate techniques, viz., path analys canonical correlation, multivariate ANOVA, multidimensional scaling, and latent structure analysi

18.1 CHARACTERISTICS AND APPLICATIONS

Multivariate techniques are largely empirical and deal with the reality; they possess the ability analyse complex data. Accordingly in most of the applied and behavioural researches, we general resort to multivariate analysis techniques for realistic results. Besides being a tool for analyzing t data, multivariate techniques also help in various types of decision-making. For example, take t case of college entrance examination wherein a number of tests are administered to candidates, an the candidates scoring high total marks based on many subjects are admitted. This system, thoug apparently fair, may at times be biased in favour of some subjects with the larger standard deviation Multivariate techniques may be appropriately used in such situations for developing norms as to wh should be admitted in college. We may also cite an example from medical field. Many medica examinations such as blood pressure and cholesterol tests are administered to patients. Each of th results of such examinations has significance of its own, but it is also important to consider relationshi between different test results or results of the same tests at different occasions in order to dra proper diagnostic conclusions and to determine an appropriate therapy. Multivariate techniques ca assist us in such a situation.

The basic objective underlying multivariate techniques is to represent a collection of massiv data in a simplified way. In other words, multivariate techniques transform a mass of observation into a smaller number of composite scores in such a way that they may reflect as much informatio as possible contained in the raw data obtained concerning a research study. Thus, the main contributio of these techniques is in arranging a large amount of complex information involved in the real dat into a simplified visible form.

For better appreciation and understanding of multivariate techniques, one must be familiar wit fundamental concepts of linear algebra, vector spaces, orthogonal and oblique projections and univariat analysis. Even then before applying multivariate techniques for meaningful results, one must conside the nature and structure of the data and the real aim of the analysis. We should also not forget tha multivariate techniques do involve several complex mathematical computations and as such can b utilized largely with the availability of computer facility.

18.2 CLASSIFICATION OF MULTIVARIATE TECHNIQUES

Today, there exist a great variety of multivariate techniques which can be conveniently classified int two broad categories viz., dependence methods and interdependence methods. This sort o classification depends upon the question: Are some of the involved variables dependent upon others If the answer is 'yes', we have dependence methods; but in case the answer is 'no', we hav interdependence methods. Two more questions are relevant for understanding the nature o multivariate techniques. Firstly, in case some variables are dependent, the question is how man variables are dependent? The other question is, whether the data are metric or non-metric? This means whether the data are quantitative, collected on interval or ratio scale, or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for a given situatio depends upon the answers to all these very questions. Jadish N. Sheth in his article on "The multivariate revolution in marketing research" has given the flow chart that clearly exhibits the nature of some important multivariate techniques as shown in Fig. 18.1.

Thus, we have two types of multivariate techniques: one type for data containing both dependent nd independent variables, and the other type for data containing several variables without dependency elationship. In the former category are included techniques like multiple regression analysis, multiple iscriminant analysis, multivariate analysis of variance and canonical analysis, whereas in the latter ategory we put techniques like factor analysis, cluster analysis, multidimensional scaling or MDS oth metric and non-metric) and the latent structure analysis.

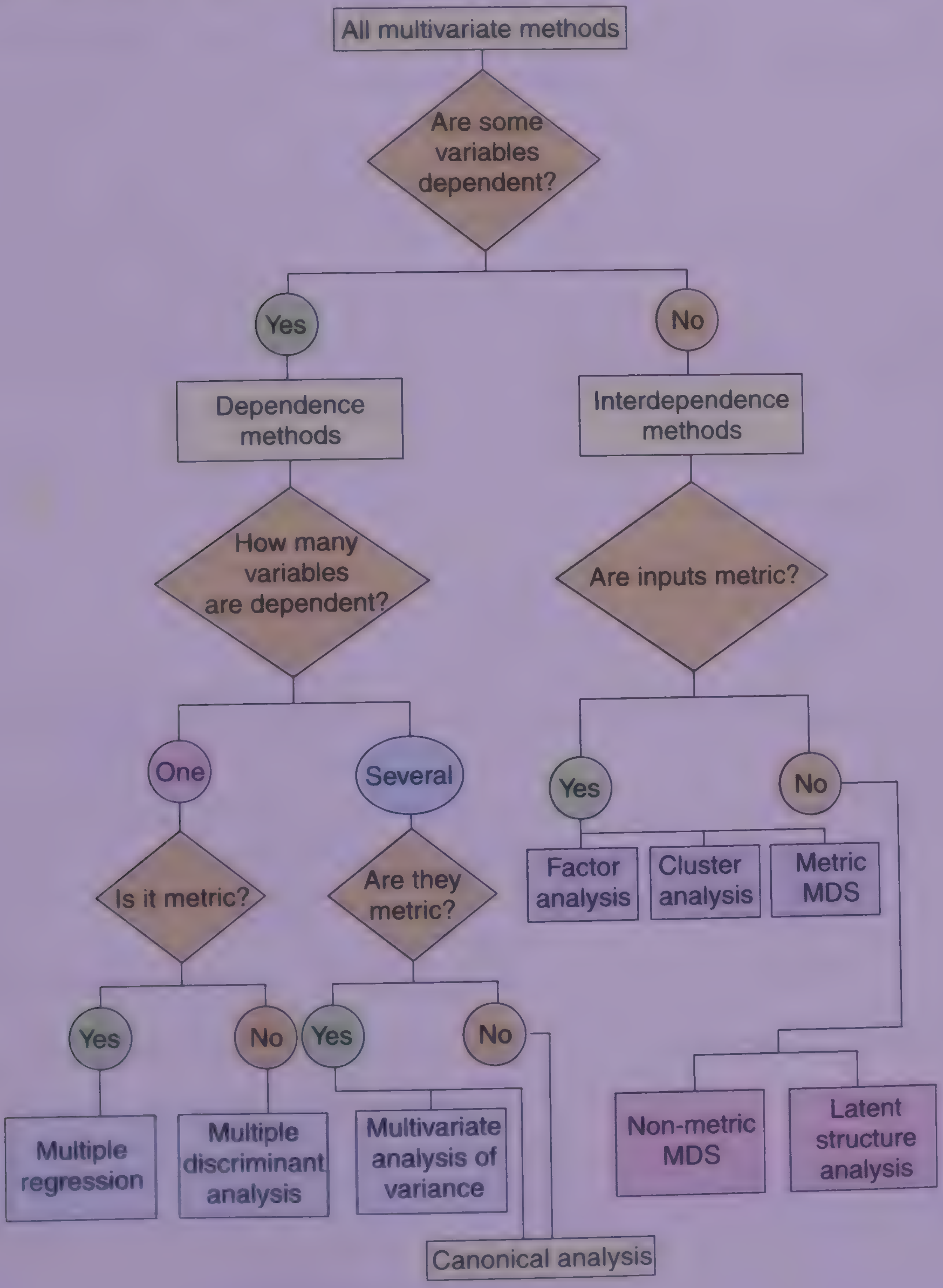

All multivariate methods
Are some variables dependent?
Yes
No
Dependence methods
Interdependence methods
How many variables are dependent?
Are inputs metric?
One
Several
Yes
No
Is it metric?
Are they metric?
Factor analysis
Cluster analysis
Metric MDS
Yes
No
Yes
No
Multiple regression
Multiple discriminant analysis
Multivariate analysis of variance
Non-metric MDS
Latent structure analysis
Canonical analysis

Fig. 18.1

18.3 SOME MULTIVARIATE TECHNIQUES

We below present brief introduction to some multivariate techniques which are not discussed in previous chapters:

18.3.1 Path Analysis

The term 'path analysis' was first introduced by the biologist Sewall Wright in 1934 in connection with decomposing the total correlation between any two variables in a causal system. The technique of path analysis is based on a series of multiple regression analyses with the added assumption of causal relationship between independent and dependent variables. This technique lays relatively heavier emphasis on the heuristic use of visual diagram, technically described as a path diagram. An illustrative path diagram showing interrelationships between Fathers' education, Fathers' occupation, Sons' education, Sons' first and Sons' present occupation can be shown in the Fig. 18.2.

Path analysis makes use of standardized partial regression coefficients (known as beta weights) as effect coefficients. In linear additive effects are assumed, then through path analysis a simple set of equations can be built up showing how each variable depends on preceding variables.

The merit of path analysis in comparison to correlational analysis is that it makes possible the assessment of the relative influence of each antecedent or explanatory variable on the consequent or criterion variables by first making explicit the assumptions underlying the causal connections and then by elucidating the indirect effect of the explanatory variables.

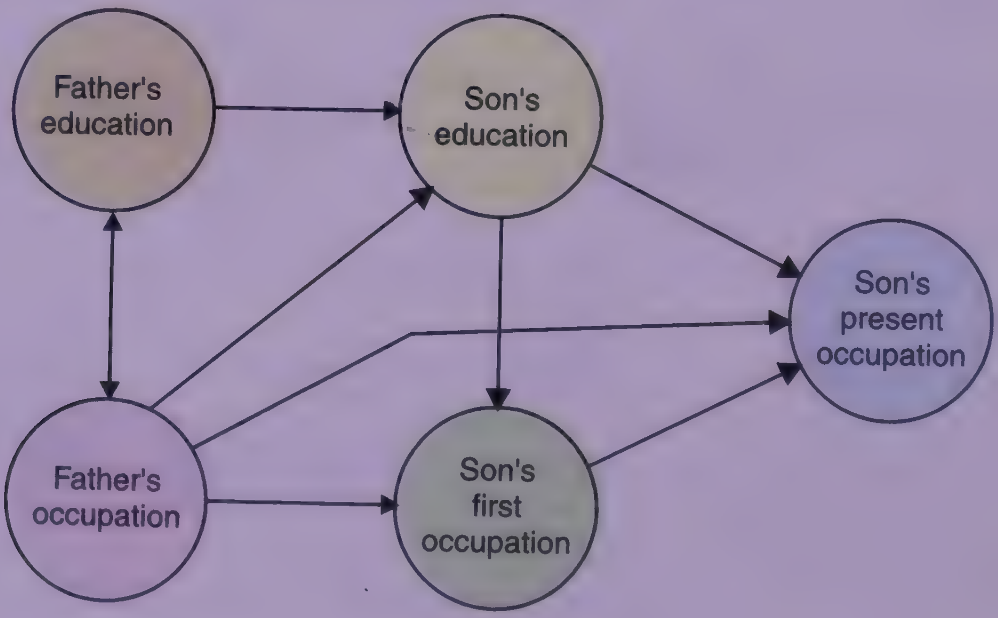


Fig.18.2

We may illustrate the path analysis technique in connection with a simple problem of testing a causal model with three explicit variables as shown in the following path diagram:

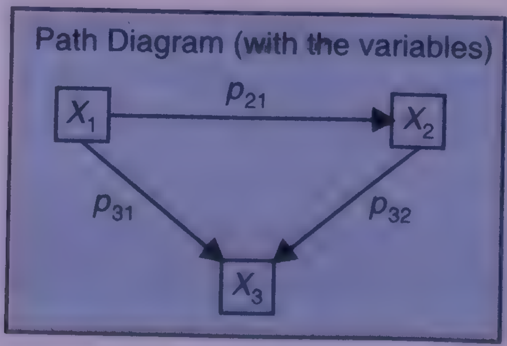


Fig. 18.3

The structural equation for the above can be written as:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \begin{matrix} = \\ = \\ = \end{matrix} \begin{bmatrix} e_1 \\ p_{21}X_1 + e_2 \\ p_{31}X_2 + p_{32}X_2 + e_3 \end{bmatrix} = pX + e$$

where the X variables are measured as deviations from their respective means. p_{21} may be estimated from the simple regression of X_2 on X_1 i.e., $X_2 = b_{21}X_1$ and p_{31} and p_{32} may be estimated from the regression of X_3 on X_2 and X_1 as under:

$$\hat{X}_3 = b_{31.2}\, X_1 + b_{2.1}\, X_2$$

where $b_{31.2}$ means the standardized partial regression coefficient for predicting variable 3 from variable 1 when the effect of variable 2 is held constant.

In path analysis the beta coefficient indicates the direct effect of X_j ($j = 1, 2, 3, ..., p$) on the dependent variable. Squaring the direct effect yields the proportion of the variance in the dependent variable Y which is due to each of the p number of independent variables X_j ($i = 1, 2, 3, ..., p$). After calculating the direct effect, one may then obtain a summary measure of the total indirect effect of X_j on the dependent variable Y by subtracting from the zero correlation coefficient r_{yxj}, the beta coefficient b_j i.e.,

$$\text{Indirect effect of } X_j \text{ on } Y = c_{jy} = r_{yxj} - b_j$$
$$\text{for all } j = 1, 2, ..., p.$$

Such indirect effects include the unanalysed effects and spurious relationships due to antecedent variables.

In the end, it may again be emphasised that the main virtue of path analysis lies in making explicit the assumptions underlying the causal connections and in elucidating the indirect effects due to antecedent variables of the given system.

18.3.2 Canonical Correlation

This technique was first developed by Hotelling wherein an effort is made to simultaneously predict a set of criterion variables from their joint co-variance with a set of explanatory variables. Both metric and non-metric data can be used in the context of this multivariate technique. The procedure followed is to obtain a set of weights for the dependent and independent variables in such a way that linear composite of the criterion variables has a maximum correlation with the linear composite of the explanatory variables. For example, if we want to relate grade school adjustment to health and physical maturity of the child, we can then use canonical correlation analysis provided, we have for each child a number of adjustment scores (such as tests, teacher's ratings, parent's ratings and so on) and also we have for each child a number of health and physical maturity scores (such as heart rate, height, weight, index of intensity of illness and so on). The main objective of canonical correlation analysis is to discover factors separately in the two sets of variables such that the multiple correlation between sets of factors will be the maximum possible. Mathematically, in canonical correlatio analysis, the weights of the two sets viz., $a_1, a_2, \ldots a_k$ and $y_1, y_2, y_3, \ldots y_j$ are so determined that th variables $X = a_1X_1 + a_2X_2 + \ldots + a_kX_k + a$ and $Y = y_1Y_1 + y_2Y_2 + \ldots y_jY_j + y$ have a maximum common variance. The process of finding the weights requires factor analyses with two matrices The resulting canonical correlation solution then gives an over all description of the presence o absence of a relationship between the two sets of variables.

18.3.3 Multidimensional Scaling

Multidimensional scaling (MDS) allows a researcher to measure an item in more than one dimension at a time. The basic assumption is that people perceive a set of objects as being more or less similar to one another on a number of dimensions (usually uncorrelated with one another) instead of only one.

There are several MDS techniques (also known as techniques for dimensional reduction) often used for the purpose of revealing patterns of one sort or another in interdependent data structures. If data happen to be non-metric, MDS involves rank ordering each pair of objects in terms of similarity. Then the judged similarities are transformed into distances through statistical manipulations and are consequently shown in *n*-dimensional space in a way that the interpoint distances best preserve the original interpoint proximities. After this sort of mapping is performed, the dimensions are usually interpreted and labeled by the researcher.

The significance of MDS lies in the fact that it enables the researcher to study "The perceptual structure of a set of stimuli and the cognitive processes underlying the development of this structure.... MDS provides a mechanism for determining the truly salient attributes without forcing the judge to appear irrational." With MDS, one can scale objects, individuals or both with a minimum of information. The MDS analysis will reveal the most salient attributes which happen to be the primary determinants for making a specific decision.

18.3.4 Multivariate ANOVA

Multivariate analysis of variance is an extension of bivariate analysis of variance in which the ratio of among-groups variance to within-groups variance is calculated on a set of variables instead of a single variable. This technique is considered appropriate when several metric dependent variables are involved in a research study along with many non-metric explanatory variables. (But if the study has only one metric dependent variable and several non-metric explanatory variables, then we use the ANOVA technique as explained earlier in the book.) In other words, multivariate analysis of variance is specially applied whenever the researcher wants to test hypotheses concerning multivariate differences in group responses to experimental manipulations. For instance, the market researcher may be interested in using one test market and one control market to examine the effect of an advertising campaign on sales as well as awareness, knowledge and attitudes. In that case he should use the technique of multivariate analysis of variance for meeting his objective.

8.3.5 Latent Structure Analysis

is type of analysis shares both of the objectives of factor analysis viz., to extract latent factors and press relationship of observed (manifest) variables with these factors as their indicators and to assify a population of respondents into pure types. This type of analysis is appropriate when the riables involved in a study do not possess dependency relationship and happen to be non-metric.

In addition to the above stated multivariate techniques, we may also describe the salient features what is known as "Path analysis", a technique useful for decomposing the total correlation between y two variables in a causal system.

18.4 CONCLUSIONS

ur aim by giving a brief introduction of some multivariate techniques in this chapter is to motivate e reader to learn these techniques in details. We refer the motivated learners to Johnson and ichern (2002). From the brief account of multivariate techniques presented above, we may conclude at such techniques are important for they make it possible to encompass all the data from an vestigation in one analysis. They in fact result in a clearer and better account of the research effort an do the piecemeal analyses of portions of data. These techniques yield more realistic probability atements in hypothesis testing and interval estimation studies. Multivariate analysis (consequently e use of multivariate techniques) is specially important in behavioural sciences and applied researches r most of such studies involve problems in which several response variables are observed multaneously. The common source of each individual observation generally results into dependence correlation among the dimensions and it is this feature that distinguishes multivariate data and chniques from their univariate prototypes.

In spite of all this, multivariate techniques are expensive and involve laborious computations. As ich their applications in the context of research studies have been accelerated only with the advent f high speed electronic computers since 1950's.

PROBLEMS

1. What do you understand by multivariate techniques?
2. Write a short note on the following multivariate techniques:
   (a) Path analysis
   (b) Multidimensional scaling
   (c) Canonical correlation
3. Write an essay on the usage of multivariate analysis in marketing research/research methodology.
4. Write an essay on the role of computers in applying various multivariate techniques.
5. How will you classify various multivariate techniques? Give details.

19

Chapter

Interpretation and Report Writing

After collecting and analyzing the data, the researcher has to accomplish the task of drawing inferenc followed by report writing. This has to be done very carefully, otherwise misleading conclusions ma be drawn and the whole purpose of doing research may get vitiated. It is only through interpretatio that the researcher can expose relations and processes that underlie his findings. In case of hypothese testing studies, if hypotheses are tested and upheld several times, the researcher may arrive a generalizations. But in case the researcher had no hypothesis to start with, he would try to explain h findings on the basis of some theory. This may at times result in new questions, leading to furthe researches. All this analytical information and consequential inference(s) may well be communicate preferably through research report, to the consumers of research results who may be either a individual or a group of individuals or some public/private organisation.

19.1 MEANING OF INTERPRETATION

Interpretation refers to the task of drawing inferences from the collected facts after an analytica and/or experimental study. In fact, it is a search for broader meaning of research findings. The tas of interpretation has two major aspects viz., (i) the effort to establish continuity in research throug linking the results of a given study with those of another, and (ii) the establishment of some explanator concepts. Interpretation is concerned with relationships within the collected data and the extensio of study beyond the collected data as well. Thus, interpretation is the device through which th factors that seem to explain what has been observed by researcher in the course of the study can b better understood and it also provides a theoretical conception which can serve as a guide for furthe researches.

Interpretation is essential for the simple reason that the usefulness and utility of research findings li in proper interpretation. It is being considered a basic component of research process because of the following reasons:

(i) It is through interpretation that the researcher can well understand the abstract principle that works beneath his findings. Through this he can link up his findings with those of other studies, having the same abstract principle, and thereby can predict about the concrete world of events. Fresh inquiries can test these predictions later on. This way the continuity in research can be maintained.

(ii) Interpretation leads to the establishment of explanatory concepts that can serve as a guide for future research studies; it opens new avenues of intellectual adventure and stimulates the quest for more knowledge.

(iii) Researcher can better appreciate only through interpretation why his findings are what they are and can make others to understand the real significance of his research findings.

(iv) The interpretation of the findings of exploratory research study often results into hypotheses for experimental research and as such interpretation is involved in the transition from exploratory to experimental research. Since an exploratory study does not have a hypothesis to start with, the findings of such a study have to be interpreted on a *post-factum* basis in which case the interpretation is technically described as '*post factum*' interpretation.

9.2 TECHNIQUES OF INTERPRETATION

ne task of interpretation is not an easy job, rather it requires a great skill and dexterity on the part of searcher. Interpretation is an art that one learns through practice and experience. The researcher ay, at times, seek the guidance from experts for accomplishing the task of interpretation.

The techniques of interpretation often involves the following steps:

(i) Researcher must give reasonable explanations of the relations which he has found and he must interpret the lines of relationship in terms of the underlying processes and must try to find out the thread of uniformity that lies under the surface layer of his diversified research findings. In fact, this is the technique of how generalization should be done and concepts be formulated.

(ii) Extraneous information, if collected during the study, must be considered while interpreting the final results of research study, for it may prove to be a key factor in understanding the problem under consideration.

(iii) It is advisable, before embarking upon final interpretation, to consult someone having insight into the study and who is frank and honest and will not hesitate to point out omissions and errors in logical argumentation. Such a consultation will result in correct interpretation and, thus, will enhance the utility of research results.

(iv) Researcher must accomplish the task of interpretation only after considering all relevant factors affecting the problem to avoid false generalization. He must be in no hurry while interpreting results, for quite often the conclusions, which appear to be all right at the beginning, may not at all be accurate.

19.3 PRECAUTIONS IN INTERPRETATION

ne should always remember that even if the data are properly collected and analysed, wrong terpretation would lead to inaccurate conclusions. It is, therefore, absolutely essential that the task of interpretation be accomplished with patience in an impartial manner and also in correct perspectiv Researcher must pay attention to the following points for correct interpretation:

(i) At the outset, researcher must invariably satisfy himself that (a) the data are appropriat trustworthy and adequate for drawing inferences; (b) the data reflect good homogeneit and that (c) proper analysis has been done through statistical methods.

(ii) The researcher must remain cautious about the errors that can possibly arise in the proce of interpreting results. Errors can arise due to false generalization and/or due to wron interpretation of statistical measures, such as the application of findings beyond the rang of observations, identification of correlation with causation and the like. Another maj pitfall is the tendency to affirm that definite relationships exist on the basis of confirmati of particular hypotheses. In fact, the positive test results accepting the hypothesis must b interpreted as "being in accord" with the hypothesis, rather than as "confirming the validit of the hypothesis". The researcher must remain vigilant about all such things so that fals generalization may not take place. He should be well equipped with and must know th correct use of statistical measures for drawing inferences concerning his study.

(iii) He must always keep in view that the task of interpretation is very much intertwined wit analysis and cannot be distinctly separated. As such he must take the task of interpretatio as a special aspect of analysis and accordingly must take all those precautions that on usually observes while going through the process of analysis viz., precautions concernin the reliability of data, computational checks, validation and comparison of results.

(iv) He must never lose sight of the fact that his task is not only to make sensitive observation of relevant occurrences, but also to identify and disengage the factors that are initiall hidden to the eye. This will enable him to do his job of interpretation on proper lines. Broa generalisation should be avoided as most research is not amenable to it because the coverag may be restricted to a particular time, a particular area and particular conditions. Suc restrictions, if any, must invariably be specified and the results must be framed within thei limits.

(v) The researcher must remember that in the course of a research study, there should b constant interaction between initial hypothesis, empirical observation and theoretica conceptions. For better interpretation of findings, the interaction between theoretica orientation and empirical observation is important.

19.4 SIGNIFICANCE OF REPORT WRITING

Research report is considered a major component of the research study for the research task remains incomplete till the report has been presented and/or written. As a matter of fact even the most brilliant hypothesis, highly well designed and conducted research study, and the most striking generalizations and findings are of little value unless they are effectively communicated to others. The purpose of research is not well served unless the findings are made known to others. Research results must invariably enter the general store of knowledge. All this explains the significance of writing research report. There are people who do not consider writing of report as an integral part of the research process. But the general opinion is in favour of treating the presentation of research results or the writing of report as part and parcel of the research project. Writing of report is the last ep in a research study and requires a set of skills somewhat different from those called for in spect of the earlier stages of research. This task should be accomplished by the researcher with most care; he may seek the assistance and guidance of experts for the purpose.

19.5 DIFFERENT STEPS IN WRITING REPORT

esearch reports are the product of slow, painstaking, accurate inductive work. The usual steps volved in writing report are: (a) logical analysis of the subject-matter; (b) preparation of the final utline; (c) preparation of the rough draft; (d) rewriting and polishing; (c) preparation of the final bliography; and (f) writing the final draft. Though all these steps are self explanatory, yet a brief ention of each one of these will be appropriate for better understanding.

(a) *Logical analysis of the subject matter:* It is the first step which is primarily concerned with the development of a subject. There are two ways in which to develop a subject (a) logically and (b) chronologically. The logical development is made on the basis of mental connections and associations between the one thing and another by means of analysis. Logical treatment often consists in developing the material from the simple possible to the most complex structures. Chronological development is based on a connection or sequence in time or occurrence. The directions for doing or making something usually follow the chronological order.

(b) *Preparation of the final outline:* Outlines are the framework upon which long written works are constructed. They are an aid to the logical organisation of the material and a reminder of the points to be stressed in the report.

(c) *Preparation of the rough draft:* This follows the logical analysis of the subject and the preparation of the final outline. Such a step is of utmost importance for the researcher now sits to write down what he has done in the context of his research study. He will write down the procedure adopted by him in collecting the material for his study along with various limitations faced by him, the technique of analysis adopted by him, the broad findings and generalizations and the various suggestions he wants to offer regarding the problem concerned.

(d) *Rewriting and polishing of the rough draft:* Usually this step requires more time than the writing of the rough draft. The careful revision makes the difference between a mediocre and a good piece of writing. While rewriting and polishing, one should check the report for weaknesses in logical development or presentation. The researcher should also (*i*) see whether or not the material, as it is presented, has unity and cohesion (ii) the report stand upright and firm and exhibit a definite pattern. In addition the researcher should give due attention to the fact that in his rough draft he has been consistent or not. He should check the mechanics of writing—grammar, spelling and usage.

(e) *Preparation of the final bibliography:* Next in order comes the task of the preparation of the final bibliography. The bibliography, which is generally appended to the research report, is a list of books in some way pertinent to the research which has been done. It should contain all those works which the researcher has consulted. The bibliography should be arranged alphabetically and may be divided into two parts; the first part may contain the names of books and pamphlets, and the second part may contain the names of magazi and newspaper articles. Generally, this pattern of bibliography is considered convenie and satisfactory from the point of view of reader, though it is not the only way of presenti bibliography. The entries in bibliography should be made adopting the following order:

For books and pamphlets the order may be as under:

(i) Name of author, last name first.
(ii) Title, underlined to indicate italics.
(iii) Publisher, place, and date of publication.
(iv) Number of volumes.

Example:

Kothari, C.R., *Quantitative Techniques,* Vikas Publishing House Pvt. Ltd., New Delhi, 1978.

For magazines and newspapers the order may be as under:

(i) Name of the author, last name first.
(ii) Title of article, in quotation marks.
(iii) Name of periodical, underlined to indicate italics.
(iv) The volume number and issue number.
(v) The date of the issue.
(vi) The pagination (page or page range as required).

Example:

Robert V. Roosa, “Coping with Short-term International Money Flows”, *The Banker,* Londo September, 1971, p. 995.

The above examples are just the samples for bibliography entries and may be used, but o should also remember that they are not the only acceptable forms. The only thing important is th whatever method one selects, it must remain consistent.

(f) *Writing the final draft:* This constitutes the last step. The final draft should be written a concise and objective style and in simple language, avoiding vague expressions such as “ seems”, “there may be”, and the like ones. While writing the final draft, the research must avoid abstract terminology and technical jargon. Illustrations and examples based o common experiences must be incorporated in the final draft as they happen to be mo effective in communicating the research findings to others. A research report should not b dull, but must enthuse people and maintain interest and must show originality. It must b remembered that every report should be an attempt to solve some intellectual problem an must contribute to the solution of a problem and must add to the knowledge of both th researcher and the reader.

19.6 LAYOUT OF THE RESEARCH REPORT

Anybody, who is reading the research report, must necessarily be conveyed enough about the stud so that he can place it in its general scientific context, judge the adequacy of its methods and thu form an opinion of how seriously the findings are to be taken. For this purpose there is the need oper layout of the report. The layout of the report means as to what the research report should ntain. A comprehensive layout of the research report should comprise (a) preliminary pages;) the main text; and (c) the end matter. Let us deal with them separately.

) Preliminary Pages

its preliminary pages the report should carry a *title and date,* followed by acknowledgements in e form of 'Preface' or 'Foreword'. Then there should be a *table of contents* followed by *list of bles and illustrations* so that the decision-maker or anybody interested in reading the report can sily locate the required information in the report.

) Main Text

e main text provides the complete outline of the research report along with all details. Title of the search study is repeated at the top of the first page of the main text and then follows the other tails on pages numbered consecutively, beginning with the second page. Each main section of the port should begin on a new page. The main text of the report should have the following sections: Introduction; (ii) Statement of findings and recommendations; (iii) The results; (iv) The implications awn from the results; and (v) The summary.

(i) *Introduction:* The purpose of introduction is to introduce the research project to the readers. It should contain a clear statement of the objectives of research i.e., enough background should be given to make clear to the reader why the problem was considered worth investigating. A brief summary of other relevant research may also be stated so that the present study can be seen in that context. The hypotheses of study, if any, and the definitions of the major concepts employed in the study should be explicitly stated in the introduction of the report.

The methodology adopted in conducting the study must be fully explained. The scientific reader would like to know in detail about such thing: How was the study carried out? What was its basic design? If the study was an experimental one, then what were the experimental manipulations? If the data were collected by means of questionnaires or interviews, then exactly what questions were asked (The questionnaire or interview schedule is usually given in an appendix)? If measurements were based on observation, then what instructions were given to the observers? Regarding the sample used in the study the reader should be told: Who were the subjects? How many were there? How were they selected? All these questions are crucial for estimating the probable limits of generalizability of the findings. The statistical analysis adopted must also be clearly stated. In addition to all this, the scope of the study should be stated and the boundary lines be demarcated. The various limitations, under which the research project was completed, must also be narrated.

(ii) *Statement of findings and recommendations:* After introduction, the research report must contain a statement of findings and recommendations in non-technical language so that it can be easily understood by all concerned. If the findings happen to be extensive, at this point they should be put in the summarised form.

(iii) *Results:* A detailed presentation of the findings of the study, with supporting data in the form of tables and charts together with a validation of results, is the next step in writing the main text of the report. This generally comprises the main body of the report, extendi over several chapters. The result section of the report should contain statistical summari and reductions of the data rather than the raw data. All the results should be presented logical sequence and splitted into readily identifiable sections. All relevant results must fi a place in the report. But how one is to decide about what is relevant is the basic questio Quite often guidance comes primarily from the research problem and from the hypothes if any, with which the study was concerned. But ultimately the researcher must rely on h own judgement in deciding the outline of his report.

(iv) *Implications of the results:* Toward the end of the main text, the researcher should aga put down the results of his research clearly and precisely. He should, state the implicatio that flow from the results of the study, for the general reader is interested in the implicatio for understanding the human behaviour. Such implications may have three aspects as stat below:

(a) A statement of the inferences drawn from the present study which may be expect to apply in similar circumstances.

(b) The conditions of the present study which may limit the extent of legitima generalizations of the inferences drawn from the study.

(c) The relevant questions that still remain unanswered or new questions raised by th study along with suggestions for the kind of research that would provide answers fo them.

It is considered a good practice to finish the report with a short conclusion which summaris and recapitulates the main points of the study. The conclusion drawn from the study shou be clearly related to the hypotheses that were stated in the introductory section. At th same time, a forecast of the probable future of the subject and an indication of the kind o research which needs to be done in that particular field is useful and desirable.

(v) *Summary:* It has become customary to conclude the research report with a very bri summary, resting in brief the research problem, the methodology, the major findings and th major conclusions drawn from the research results.

(c) End Matter

At the end of the report, appendices should be enlisted in respect of all technical data such a questionnaires, sample information, mathematical derivations and the like ones. Bibliography of source consulted should also be given. Index (an alphabetical listing of names, places and topics along wit the numbers of the pages in a book or report on which they are mentioned or discussed) shoul invariably be given at the end of the report. The value of index lies in the fact that it works as a guid to the reader for the contents in the report.

19.7 TYPES OF REPORTS

Research reports vary greatly in length and type. In each individual case, both the length and th form are largely dictated by the problems at hand. For instance, business firms prefer reports in th letter form, just one or two pages in length. Banks, insurance organisations and financial institution are generally fond of the short balance-sheet type of tabulation for their annual reports to the customers and shareholders. Mathematicians prefer to write the results of their investigations in th form of algebraic notations. Chemists report their results in symbols and formulae. Students o iterature usually write long reports presenting the critical analysis of some writer or period or the like vith a liberal use of quotations from the works of the author under discussion. In the field of education nd psychology, the favourite form is the report on the results of experimentation accompanied by he detailed statistical tabulations. Clinical psychologists and social pathologists frequently find it ecessary to make use of the case-history form.

News items in the daily papers are also forms of report writing. They represent firsthand on-the-cene accounts of the events described or compilations of interviews with persons who were on the cene. In such reports the first paragraph usually contains the important information in detail and the ucceeding paragraphs contain material which is progressively less and less important.

Book-reviews which analyze the content of the book and report on the author's intentions, his uccess or failure in achieving his aims, his language, his style, scholarship, bias or his point of view. Such reviews also happen to be a kind of short report. The reports prepared by governmental bureaus, pecial commissions, and similar other organisations are generally very comprehensive reports on he issues involved. Such reports are usually considered as important research products. Similarly, Ph.D. theses and dissertations are also a form of report-writing, usually completed by students in cademic institutions.

The above narration throws light on the fact that the results of a research investigation can be resented in a number of ways viz., a technical report, a popular report, an article, a monograph or at imes even in the form of oral presentation. Which method(s) of presentation to be used in a particular tudy depends on the circumstances under which the study arose and the nature of the results. A *echnical report* is used whenever a full written report of the study is required whether for record-eeping or for public dissemination. A *popular report* is used if the research results have policy mplications. We give below a few details about the said two types of reports:

a) Technical Report

n the technical report the main emphasis is on (i) the methods employed, (ii) assumptions made in he course of the study, (iii) the detailed presentation of the findings including their limitations and upporting data.

A general outline of a technical report can be as follows:

(i) *Summary of results:* A brief review of the main findings just in two or three pages.

(ii) *Nature of the study:* Description of the general objectives of study, formulation of the problem in operational terms, the working hypothesis, the type of analysis and data required, etc.

(iii) *Methods employed:* Specific methods used in the study and their limitations. For instance, in sampling studies we should give details of sample design viz., sample size, sample selection, etc.

(iv) *Data:* Discussion of data collected, their sources, characteristics and limitations. If secondary data are used, their suitability to the problem at hand be fully assessed. In case of a survey, the manner in which data were collected should be fully described.

(v) *Analysis of data and presentation of findings:* The analysis of data and presentation of the findings of the study with supporting data in the form of tables and charts be fully narrated. This, in fact, happens to be the main body of the report usually extending over several chapters.

(vi) *Conclusions:* A detailed summary of the findings and the policy implications drawn from the results be explained.

(vii) *Bibliography:* Bibliography of various sources consulted be prepared and attached.

(viii) *Technical appendices:* Appendices be given for all technical matters relating to questionnaire, mathematical derivations, elaboration on particular technique of analysis and the like ones.

(ix) *Index:* Index must be prepared and be given invariably in the report at the end.

The order presented above only gives a general idea of the nature of a technical report; the order of presentation may not necessarily be the same in all the technical reports. This, in other words, means that the presentation may vary in different reports; even the different sections outlined above will not always be the same, nor will all these sections appear in any particular report.

It should, however, be remembered that even in a technical report, simple presentation and ready availability of the findings remain an important consideration and as such the liberal use of charts and diagrams is considered desirable.

(b) Popular Report

The popular report is one which gives emphasis on simplicity and attractiveness. The simplification should be sought through clear writing, minimization of technical, particularly mathematical, details and liberal use of charts and diagrams. Attractive layout along with large print, many subheadings, even an occasional cartoon now and then is another characteristic feature of the popular report. Besides, in such a report emphasis is given on practical aspects and policy implications.

We give below a general outline of a popular report.

(i) *The findings and their implications:* Emphasis in the report is given on the findings of most practical interest and on the implications of these findings.

(ii) *Recommendations for action:* Recommendations for action on the basis of the findings of the study is made in this section of the report.

(iii) *Objective of the study:* A general review of how the problem arise is presented along with the specific objectives of the project under study.

(iv) *Methods employed:* A brief and non-technical description of the methods and techniques used, including a short review of the data on which the study is based, is given in this part of the report.

(v) *Results:* This section constitutes the main body of the report wherein the results of the study are presented in clear and non-technical terms with liberal use of all sorts of illustrations such as charts, diagrams and the like ones.

(vi) *Technical appendices:* More detailed information on methods used, forms, etc. is presented in the form of appendices. But the appendices are often not detailed if the report is entirely meant for general public.

There can be several variations of the form in which a popular report can be prepared. The only important thing about such a report is that it gives emphasis on simplicity and policy implications from the operational point of view, avoiding the technical details of all sorts to the extent possible.

19.8 ORAL PRESENTATION

t times oral presentation of the results of the study is considered effective, particularly in cases here policy recommendations are indicated by project results. The merit of this approach lies in the act that it provides an opportunity for give-and-take decisions which generally lead to a better nderstanding of the findings and their implications. But the main demerit of this sort of presentation s the lack of any permanent record concerning the research details and it may be just possible that he findings may fade away from people's memory even before an action is taken. In order to vercome this difficulty, a written report may be circulated before the oral presentation and referred o frequently during the discussion. Oral presentation is effective when supplemented by various isual devices. Use of slides, wall charts and blackboards is quite helpful in contributing to clarity and n reducing the boredom, if any. Distributing a board outline, with a few important tables and charts oncerning the research results, makes the listeners attentive who have a ready outline on which to ocus their thinking. This very often happens in academic institutions where the researcher discusses is research findings and policy implications with others either in a seminar or in a group discussion.

Thus, research results can be reported in more than one ways, but the usual practice adopted, in cademic institutions particularly, is that of writing the Technical Report and then preparing several esearch papers to be discussed at various forums in one form or the other. But in practical field and vith problems having policy implications, the technique followed is that of writing a popular report. Researches done on governmental account or on behalf of some major public or private organisations re usually presented in the form of technical reports.

19.9 MECHANICS OF WRITING A RESEARCH REPORT

There are very definite and set rules which should be followed in the actual preparation of the esearch report or paper. Once the techniques are finally decided, they should be scrupulously adhered to, and no deviation permitted. The criteria of format should be decided as soon as the materials for the research paper have been assembled. The following points deserve mention so far as the mechanics of writing a report are concerned:

(i) *Size and physical design:* The manuscript should be written on unruled paper $8\ 1/2'' \times 11''$ in size. If it is to be written by hand, then black or blue-black ink should be used. A margin of at least one and one-half inches should be allowed at the left hand and of at least half an inch at the right hand of the paper. There should also be one-inch margins, top and bottom. The paper should be neat and legible. If the manuscript is to be typed, then all typing should be double-spaced on one side of the page only except for the insertion of the long quotations.

(ii) *Procedure:* Various steps in writing the report should be strictly adhered (All such steps have already been explained earlier in this chapter).

(iii) *Layout:* Keeping in view the objective and nature of the problem, the layout of the report should be thought of and decided and accordingly adopted (The layout of the research report and various types of reports have been described in this chapter earlier which should be taken as a guide for report-writing in case of a particular problem).

(iv) *Treatment of quotations:* Quotations should be placed in quotation marks and doubl spaced, forming an immediate part of the text. But if a quotation is of a considerable lengt (more than four or five type written lines) then it should be single-spaced and indented a least half an inch to the right of the normal text margin.

(v) *The footnotes:* Regarding footnotes one should keep in view the following:

(a) The footnotes serve two purposes viz., the identification of materials used in quotation in the report and the notice of materials not immediately necessary to the body of th research text but still of supplemental value. In other words, footnotes are meant fo cross references, citation of authorities and sources, acknowledgement and elucidatio or explanation of a point of view. It should always be kept in view that footnote is no an end nor a means of the display of scholarship. The modern tendency is to make th minimum use of footnotes for scholarship does not need to be displayed.

(b) Footnotes are placed at the bottom of the page on which the reference or quotatio which they identify or supplement ends. Footnotes are customarily separated from th textual material by a space of half an inch and a line about one and a half inches long

(c) Footnotes should be numbered consecutively, usually beginning with 1 in each chapte separately. The number should be put slightly above the line, say at the end of a quotation At the foot of the page, again, the footnote number should be indented and typed a littl above the line. Thus, consecutive numbers must be used to correlate the reference i the text with its corresponding note at the bottom of the page, except in case of statistica tables and other numerical material, where symbols such as the asterisk (*) or the lik one may be used to prevent confusion.

(d) Footnotes are always typed in single space though they are divided from one anothe by double space.

(vi) *Documentation style:* Regarding documentation, the first footnote reference to any given work should be complete in its documentation, giving all the essential facts about the edition used. Such documentary footnotes follow a general sequence. The common order may be described as under:

(A) *Regarding the single-volume reference*

(a) Author's name in normal order (and not beginning with the last name as in a bibliography followed by a comma;

(b) Title of work, underlined to indicate italics;

(c) Place and date of publication;

(d) Pagination references (The page number).

Example:

John Gassner, *Masters of the Drama,* New York: Dover Publications, Inc. 1954, p. 315.

(B) *Regarding multivolumed reference*

(a) Author's name in the normal order;

(b) Title of work, underlined to indicate italics;

(c) Place and date of publication;

(d) Number of volume;

(e) Pagination references (The page number).

C) *Regarding works arranged alphabetically*

For works arranged alphabetically such as encyclopedias and dictionaries, no pagination reference is usually needed. In such cases the order is illustrated as under:

xample 1:

"Salamanca," *Encyclopaedia Britannica,* 14th Edition.

xample 2:

"Mary Wollstonecraft Godwin," *Dictionary of national biography.*

But if there should be a detailed reference to a long encyclopedia article, volume and pagination reference may be found necessary.

D) *Regarding periodicals reference*

(a) Name of the author in normal order;

(b) Title of article, in quotation marks;

(c) Name of periodical, underlined to indicate italics;

(d) Volume number;

(e) Date of issuance;

(f) Pagination.

(E) *Regarding anthologies and collections reference*

Quotations from anthologies or collections of literary works must be acknowledged not only by author, but also by the name of the collector.

(F) *Regarding second-hand quotations reference*

In such cases the documentation should be handled as follows:

(a) Original author and title;

(b) "quoted or cited in,";

(c) Second author and work.

Example:

J.F. Jones, *Life in Ploynesia,* p. 16, quoted in *History of the Pacific Ocean area,* by R.B. Abel, p. 191.

(G) *Case of multiple authorship*

If there are more than two authors or editors, then in the documentation the name of only the first is given and the multiple authorship is indicated by "et al." or "and others".

Subsequent references to the same work need not be so detailed as stated above. If the work is cited again without any other work intervening, it may be indicated as *ibid.* followed by a comma and the page number. A single page should be referred to as p., but more than one page be referred to as pp. If there are several pages referred to at a stretch, the practice is to use often the page number.

for example, pp. 190ff, which means page number 190 and the following pages; but only for page 1 and the following page ‘190f’. Roman numerical is generally used to indicate the number of t volume of a book. Op. cit. (opera citato, in the work cited) or Loc. cit. (loco citato, in the place cite are two of the very convenient abbreviations used in the footnotes. Op. cit. or Loc. cit. after t writer’s name would suggest that the reference is to work by the writer which has been cited detail in an earlier footnote but intervened by some other references.

(vii) *Punctuation and abbreviations in footnotes:* The first item after the number in t footnote is the author’s name, given in the normal signature order. This is followed by comma. After the comma, the title of the book is given: the article (such as “A”, “An “The” etc.) is omitted and only the first word and proper nouns and adjectives are capitalize The title is followed by a comma. Information concerning the edition is given next. Th entry is followed by a comma. The place of publication is then stated; it may be mentione in an abbreviated form, if the place happens to be a famous one such as Lond. for Londo N.Y. for New York, N.D. for New Delhi and so on. This entry is followed by a comm Then the name of the publisher is mentioned and this entry is closed by a comma. It followed by the date of publication if the date is given on the title page. If the date appea in the copyright notice on the reverse side of the title page or elsewhere in the volume, th comma should be omitted and the date enclosed in square brackets [c 1978], [1978]. Th entry is followed by a comma. Then follow the volume and page references and ar separated by a comma if both are given. A period closes the complete documentary referenc But one should remember that the documentation regarding acknowledgements fro magazine articles and periodical literature follow a different form as stated earlier whil explaining the entries in the bibliography.

Certain English and Latin abbreviations are quite often used in bibliographies and footnotes t eliminate tedious repetition. The following is a partial list of the most common abbreviations frequentl used in report-writing (the researcher should learn to recognise them as well as he should learn t use them):

anon.,	anonymous
ante.,	before
art.,	article
aug.,	augmented
bk.,	book
bull.,	bulletin
cf.,	compare
ch.,	chapter
col.,	column
diss.,	dissertation
ed.,	editor, edition, edited.
ed. cit.,	edition cited
e.g.,	exempli gratia: for example
eng.,	enlarged
et.al.,	and others

et seq.,	et sequens: and the following
ex.,	example
f., ff.,	and the following
fig(s).,	figure(s)
fn.,	footnote
ibid., ibidem:	in the same place (when two or more successive footnotes refer to the same work, it is not necessary to repeat complete reference for the second footnote. Ibid. may be used. If different pages are referred to, pagination must be shown).
id., idem:	the same
ill., illus., or illust(s).	illustrated, illustration(s)
Intro., intro.,	introduction
l, or *ll*,	line(s)
loc. cit., loco citato:	in the place cited; used as op.cit., (when new reference is made to the same pagination as cited in the previous note)
MS., MSS.,	Manuscript or Manuscripts
N.B., nota bene:	note well
n.d.,	no date
n.p.,	no place
no pub.,	no publisher
no(s).,	number(s)
o.p.,	out of print
op. cit: opera citato	in the work cited (If reference has been made to a work and new reference is to be made, ibid., may be used, if intervening reference has been made to different works, op.cit. must be used. The name of the author must precede.
p. or pp.,	page(s)
passim:	here and there
post:	after
rev.,	revised
tr., trans.,	translator, translated, translation
vid or vide:	see, refer to
viz.,	namely
vol. or vol(s).,	volume(s)
vs., versus:	against

(viii) *Use of statistics, charts and graphs:* A judicious use of statistics in research reports is often considered a virtue for it contributes a great deal towards the clarification and simplification of the material and research results. One may well remember that a good picture is often worth more than a thousand words. Statistics are usually presented in the form of tables, charts, bars and line-graphs and pictograms. Such presentation should b self explanatory and complete in itself. It should be suitable and appropriate looking to th problem at hand. Finally, statistical presentation should be neat and attractive.

(ix) *The final draft:* Revising and rewriting the rough draft of the report should be done wit great care before writing the final draft. For the purpose, the researcher should put t himself questions like: Are the sentences written in the report clear? Are they grammaticall correct? Do they say what is meant'? Do the various points incorporated in the report fi together logically? "Having at least one colleague read the report just before the fina revision is extremely helpful. Sentences that seem crystal-clear to the writer may prov quite confusing to other people; a connection that had seemed self evident may strik others as a *non-sequitur.* A friendly critic, by pointing out passages that seem unclear o illogical, and perhaps suggesting ways of remedying the difficulties, can be an invaluabl aid in achieving the goal of adequate communication."*

(x) *Bibliography:* Bibliography should be prepared and appended to the research report a discussed earlier.

(xi) *Preparation of the index:* At the end of the report, an index should invariably be given, the value of which lies in the fact that it acts as a good guide, to the reader. Index may be prepared both as subject index and as author index. The former gives the names of the subject-topics or concepts along with the number of pages on which they have appeared o discussed in the report, whereas the latter gives the similar information regarding the names of authors. The index should always be arranged alphabetically. Some people prefer to prepare only one index common for names of authors, subject-topics, concepts and the like ones.

19.10 PRECAUTIONS FOR WRITING RESEARCH REPORTS

Research report is a channel of communicating the research findings to the readers of the report. A good research report is one which does this task efficiently and effectively. As such it must be prepared keeping the following precautions in view:

1. While determining the length of the report (since research reports vary greatly in length), one should keep in view the fact that it should be long enough to cover the subject but short enough to maintain interest. In fact, report-writing should not be a means to learning more and more about less and less.
2. A research report should not, if this can be avoided, be dull; it should be such as to sustain reader's interest.
3. Abstract terminology and technical jargon should be avoided in a research report. The report should be able to convey the matter as simply as possible. This, in other words, means that report should be written in an objective style in simple language, avoiding expressions such as "it seems," "there may be" and the like.

---

* Claire Selltiz and others, *Research Methods in Social Relations* rev., Methuen & Co. Ltd., London, 1959, p. 454.

4. Readers are often interested in acquiring a quick knowledge of the main findings and as such the report must provide a ready availability of the findings. For this purpose, charts, graphs and the statistical tables may be used for the various results in the main report in addition to the summary of important findings.
5. The layout of the report should be well thought out and must be appropriate and in accordance with the objective of the research problem.
6. The reports should be free from grammatical mistakes and must be prepared strictly in accordance with the techniques of composition of report-writing such as the use of quotations, footnotes, documentation, proper punctuation and use of abbreviations in footnotes and the like.
7. The report must present the logical analysis of the subject matter. It must reflect a structure wherein the different pieces of analysis relating to the research problem fit well.
8. A research report should show originality and should necessarily be an attempt to solve some intellectual problem. It must contribute to the solution of a problem and must add to the store of knowledge.
9. Towards the end, the report must also state the policy implications relating to the problem under consideration. It is usually considered desirable if the report makes a forecast of the probable future of the subject concerned and indicates the kinds of research still needs to be done in that particular field.
10. Appendices should be enlisted in respect of all the technical data in the report.
11. Bibliography of sources consulted is a must for a good report and must necessarily be given.
12. Index is also considered an essential part of a good report and as such must be prepared and appended at the end.
13. Report must be attractive in appearance, neat and clean, whether typed or printed.
14. Calculated confidence limits must be mentioned and the various constraints experienced in conducting the research study may also be stated in the report.
15. Objective of the study, the nature of the problem, the methods employed and the analysis techniques adopted must all be clearly stated in the beginning of the report in the form of introduction.

19.11 CONCLUSION

In spite of all that has been stated above, one should always keep in view the fact report-writing is an art which is learnt by practice and experience, rather than by mere doctrination.

PROBLEMS

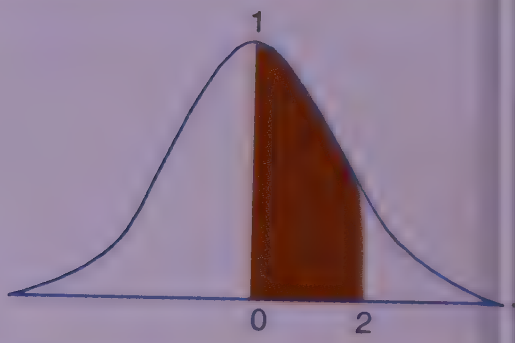
1. Write a brief note on the 'task of interpretation' in the context of research methodology.
2. "Interpretation is a fundamental component of research process". Explain. Why so?
3. Describe the precautions that the researcher should take while interpreting his findings.

4. "Interpretation is an art of drawing inferences, depending upon the skill of the researcher". Elucidate the given statement explaining the technique of interpretation.
5. "It is only through interpretation the researcher can expose the relations and processes that underlie his findings". Explain, giving examples.
6. Explain the significance of a research report and narrate the various steps involved in writing such a report.
7. Describe, in brief, the layout of a research report, covering all relevant points.
8. Write a short note on 'Documentation' in the context of a research report.
9. Mention the different types of report, particularly pointing qut the difference between a technical report and a popular report.
10. Explain the technique and importance of oral presentation of research findings. Is only oral presentation sufficient? If not, why?
11. (a) What points will you keep in mind while preparing a research report? Explain.
    (b) What are the different forms in which a research work may be reported. Describe.
12. "We can teach methods of analysis, yet any extensive research... requires something equally important: an organisation or synthesis which provides the essential structure into which the pieces of analysis fit." Examine this statement and show how a good research report may be prepared.
13. Write short notes on the following:
    (a) The techniques of writing report;
    (b) Characteristics of a good research report;
    (c) Bibliography and its importance in context of research report;
    (d) Rewriting and polishing of report.
14. "Report writing is more an art that hinges upon practice and experience". Discuss.

Appendix
(Selected Statistical Tables)

Table 1: Area Under Normal Curve

An entry in the table is the proportion under the entire curve which is between $z = 0$ and a positive value of z. Areas for negative values for z are obtained by symmetry.



Areas of a standard normal distribution

z	0	0.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2903	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table 2: Critical Values of Student's *t*-Distribution

d.f.	Level of significance for two-tailed test					d.f.
	0.20	0.10	0.05	0.02	0.01	
	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.01	0.005	
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.731	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
Infinity	1.282	1.645	1.960	2.326	2.576	Infinity

Table 3: Critical Values of χ^2

Degrees of freedom	Probability under H_0 that of χ^2 > Chi square						
	.99	.95	.50	.10	.05	.02	.01
1	.000157	.00393	.455	2.706	3.841	5.412	6.635
2	.0201	.103	1.386	4.605	5.991	7.824	9.210
3	.115	.352	2.366	6.251	7.815	9.837	11.341
4	.297	.711	3.357	7.779	9.488	11.668	13.277
5	.554	.1145	4.351	9.236	11.070	13.388	15.086
6	.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.239	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.342	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725
12	3.571	5.226	11.340	18.549	21.026	24.054	26.217
13	4.107	5.892	12.340	19.812	22.362	25.472	72.688
14	4.660	6.571	13.339	21.064	23.685	26.873	29.141
15	4.229	7.261	14.339	22.307	24.996	28.259	30.578
16	5.812	7.962	15.338	23.542	26.296	29.633	32.000
17	6.408	8.672	16.338	24.769	27.587	30.995	33.409
18	7.015	9.390	17.338	25.989	28.869	32.346	34.805
19	7.633	10.117	18.338	27.204	30.144	33.687	36.191
20	8.260	10.851	19.337	28.412	31.410	35.020	37.566
21	8.897	11.591	20.337	29.615	32.671	36.343	38.932
22	9.542	12.338	21.337	30.813	33.924	37.659	40.289
23	10.196	13.091	22.337	32.007	35.172	38.968	41.638
24	10.856	13.848	23.337	32.196	36.415	40.270	42.980
25	11.524	14.611	24.337	34.382	37.652	41.566	44.314
26	12.198	15.379	25.336	35.363	38.885	41.856	45.642
27	12.879	16.151	26.336	36.741	40.113	44.140	46.963
28	13.565	16.928	27.336	37.916	41.337	45.419	48.278
29	14.256	17.708	28.336	39.087	42.557	46.693	49.588
30	14.953	18.493	29.336	40.256	43.773	47.962	50.892

Note: For degrees of freedom greater than 30, the quantity $2\chi^2 - \sqrt{2\text{d.f.} - 1}$ may be used as a normal variate with unit variance i.e., $z_\alpha = \sqrt{2\chi^2} - \sqrt{2\text{d.f.} - 1}$.

Table 4(a): Critical Values of F-Distribution (at 5 per cent)

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.1	243.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

v_1 = Degrees of freedom for greater variance.

v_2 = Degrees of freedom for smaller variance.

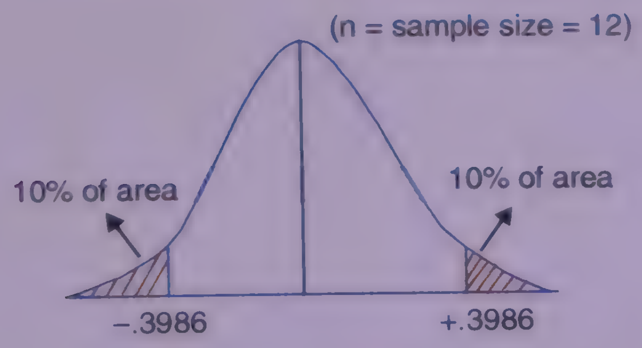
Table 4(b): Critical Values of *F*-Distribution (at 1 per cent)

v_1 / v_2	1	2	3	4	5	6	8	12	24	∞
1	4052	4999.5	5403	5625	5764	5859	5982	6106	6235	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.45
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.87	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	3.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.20	2.96	2.58	2.10
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.45	2.13
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

v_1 = Degrees of freedom for greater variance.

v_2 = Degrees of freedom for smaller variance.

Table 5: Values for Spearman's Rank Correlation (r_s) for Combined Areas in Both Tails



n	.20	.10	.05	.02	.01	.002
4	.8000	.8000	—	—	—	—
5	.7000	.8000	.9000	.9000	—	—
6	.6000	.7714	.8236	.8857	.9429	—
7	.5357	.6786	.7450	.8571	.8929	.9643
8	.5000	.6190	.7143	.8095	.8571	.9286
9	.4667	.5833	.6833	.7667	.8167	.9000
10	.4424	.5515	.6364	.7333	.7818	.8667
11	.4182	.5273	.6091	.7000	.7455	.8364
12	.3986	.4965	.5804	.6713	.7273	.8182
13	.3791	.4780	.5549	.6429	.6978	.7912
14	.3626	.4593	.5341	.6220	.6747	.7670
15	.3500	.4429	.5179	.6000	.6536	.7464
16	.3382	.4265	.5000	.5824	.6324	.7265
17	.3260	.4118	.4853	.5637	.6152	.7083
18	.3148	.3994	.4716	.5480	.5975	.6904
19	.3070	.3895	.4579	.5333	.5825	.6737
20	.2977	.3789	.4451	.5203	.5684	.6586
21	.2909	.3688	.4351	.5078	.5545	.6455
22	.2829	.3597	.4241	.4963	.5426	.6318
23	.2767	.3518	.4150	.4852	.5306	.6186
24	.2704	.3435	.4061	.4748	.5200	.6070
25	.2646	.3362	.3977	.4654	.5100	.5962
26	.2588	.3299	.3894	.4564	.5002	.5856
27	.2540	.3236	.3822	.4481	.4915	.5757
28	.2480	.3175	.3749	.4401	.4828	.5660
29	.2443	.3113	.3685	.4320	.4744	.5567
30	.2400	.3059	.3620	.4251	.4665	.5479

Table 6: Selected Values of Wilcoxon's (Unpaired) Distribution

$[W_s - \text{Min } W_s]$ or $[\text{Max. } W_l - W_l]$

s	l	Min W_s	Max W_s	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	2	3	7																					
	3	3	12	.100																				
	4	3	18	.067	.134																			
	5	3	25	.048	.095	.190																		
	6	3	33	.086	.071	.143																		
	7	3	42	.028	.056	.111																		
	8	3	52	.022	.044	.089	.133																	
3	2	6	9	.100																				
	3	6	15	.050	.100																			
	4	6	22	.029	.057	.114																		
	5	6	30	.018	.036	.071	.125																	
	6	6	39	.012	.024	.048	.083	.131																
	7	6	49	.008	.017	.033	.058	.092	.133															
	8	6	60	.006	.012	.024	.042	.067	.097	.139														
4	2	10	11	.007	*																			
	3	10	18	.029	.057	.114																		
	4	10	26	.014	.029	.057	.100																	
	5	10	35	.008	.016	.032	.056	.095	.143															
	6	10	45	.005	.010	.019	.033	.057	.086	.129														
	7	10	56	.003	.006	.012	.021	.036	.055	.082	.115													
	8	10	68	.002	.004	.008	.014	.024	.036	.055	.077	.107												
5	3	15	21	.018	.036	.071	.125	*																
	4	15	30	.008	.016	.032	.056	.095	.143															
	5	15	40	.004	.008	.016	.028	.048	.075	.111														
	6	15	51	.002	.004	.009	.015	.026	.041	.063	.089	.123												
	7	15	63	.001	.003	.005	.009	.015	.024	.037	.053	.074	.101											
	8	15	76	.001	.002	.003	.005	.009	.015	.023	.033	.047	.064	.085	.111									

(Contd.)

s	l	Min W	Max W	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	3	21	24	.012	.024	*																		
	4	21	34	.005	.010	.019	.033	.057	.086	.129	*													
	5	21	45	.002	.004	.009	.015	.026	.041	.063	.089	.123												
	6	21	57	.001	.002	.004	.008	.013	.021	.032	.047	.066	.090	.120										
	7	21	70	.001	.001	.002	.004	.007	.011	.017	.026	.037	.051	.069	.090	.117								
	8	21	84	.000	.001	.001	.002	.004	.006	.010	.015	.021	.030	.041	.054	.071	.091	.114						
7	4	28	38	.003	.006	.012	.021	.036	.055*															
	5	28	50	.001	.003	.005	.009	.015	.024	.037	.053	.074	.101											
	6	28	63	.001	.001	.002	.004	.007	.011	.017	.026	.037	.051	.069	.090	.117								
	7	28	77	.000	.001	.001	.002	.003	.006	.009	.013	.019	.027	.036	.049	.064	.082	.104						
	8	28	92	.000	.000	.001	.001	.002	.003	.005	.007	.010	.014	.020	.027	.036	.047	.060	.076	.095	.116			
8	4	36	42	.002	.004	.008	.014	*																
	5	36	55	.001	.002	.003	.005	.009	.015	.023	.033	.047	.064	*										
	6	36	69	.000	.001	.001	.002	.004	.006	.010	.015	.021	.030	.041	.054	.071	.091	.114						
	7	36	84	.000	.000	.001	.001	.002	.003	.005	.007	.010	.014	.020	.027	.036	.047	.060	.076	.095	.116			
	8	36 -	100	.000	.000	.000	.001	.001	.001	.002	.003	.005	.007	.010	.014	.019	.025	.032	.041	.052	.065	.080	.097	.117

* Indicates that the value at head of this column (add those values that are larger) are not possible for the given values of *s* and *l* in this row.

Table 7: Quantiles of the Wilcoxon Rank Test Statistic

	$w_{0.005}$	$w_{0.01}$	$w_{0.025}$	$w_{0.05}$	$w_{0.10}$	$w_{0.20}$	$w_{0.30}$	$w_{0.40}$	$w_{0.50}$	$\frac{n(n+1)}{2}$
$n =$ 4	0	0	0	0	1	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	41	48	56	66	73	80	85.5	171
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	83	91	98	105	210
21	44	50	59	68	78	91	100	108	115.5	131
22	49	56	67	76	87	100	110	119	126.5	153
23	55	63	74	84	95	110	120	130	138	176
24	62	70	82	92	105	120	131	141	150	300
25	69	77	90	101	114	131	143	153	162.5	325
26	76	85	99	111	125	142	155	165	175.5	351
27	84	94	108	120	135	154	167	178	189	378
28	92	102	117	131	146	166	180	192	203	406
29	101	111	127	141	158	178	193	206	217.5	435
30	110	121	138	152	170	191	207	220	232.5	465
31	119	131	148	164	182	205	221	235	248	496
32	129	141	160	176	195	219	236	250	264	528
33	139	152	171	188	208	233	251	266	280.5	561
34	149	163	183	201	222	248	266	282	297.5	595
35	160	175	196	214	236	263	283	299	315	630
36	172	187	209	228	251	279	299	317	333	666
37	184	199	222	242	266	295	316	335	351.5	703
38	196	212	236	257	282	312	334	353	370.5	741
39	208	225	250	272	298	329	352	372	390	780
40	221	239	265	287	314	347	371	391	410	820

(Contd.)

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	$\frac{n(n+1)}{2}$
41	235	253	280	303	331	365	390	411	430.5	861
42	248	267	295	320	349	384	409	431	451.5	903
43	263	282	311	337	366	403	429	452	473	946
44	277	297	328	354	385	422	450	473	495	990
45	292	313	344	372	403	442	471	495	517.5	1035
46	308	329	362	390	423	463	492	517	540.5	1081
47	324	346	379	408	442	484	514	540	564	1128
48	340	363	397	428	463	505	536	563	588	1176
49	357	381	416	447	483	527	559	587	612.5	1225
50	374	398	435	467	504	550	583	611	637.5	1275

For n larger than 50, the pth quantile Wp of the Wilcoxon signed ranked test statistic may be approximated by $w_p = \left[\frac{n(n+1)}{4}\right] = x_p\sqrt{\frac{n(n+1)(2n+1)}{24}}$, where x_p is the p^{th} quantile of a standard normal random variable.

Table 8: The Kruskal - Wallis test

Critical region : H ≥ Tabulated value

K = 3					K = 4						K = 5						
Sample sizes			α = 0.05	α = 0.01	Sample sizes				α = 0.05	α = 0.01	Sample sizes					α = 0.05	α = 0.01
2	2	2	–	–	2	2	1	1	–	–	2	2	1	1	1	–	–
3	2	1	–	–	2	2	2	1	5.679	–	2	2	2	1	1	6.750	–
3	2	2	4.714	–	2	2	2	2	6.167	6.667	2	2	2	2	1	7.133	7.533
3	3	1	5.143	–	3	1	1	1	–	–	2	2	2	2	2	7.418	8.291
3	3	2	5.361	–	3	2	1	1	–	–	3	1	1	1	1	–	–
3	3	3	5.600	7.200	3	2	2	1	5.833	–	3	2	1	1	1	6.583	–
4	2	1	–	–	3	2	2	2	6.333	7.133	3	2	2	1	1	6.800	7.600
4	2	2	5.333	–	3	3	1	1	6.333	–	3	2	2	2	1	7.309	8.127
4	3	1	5.208	–	3	3	2	1	6.244	7.200	3	2	2	2	2	7.682	8.682
4	3	2	5.444	6.444	3	3	2	2	6.527	7.636	3	3	1	1	1	7.111	–
4	3	3	5.791	6.745	3	3	3	1	6.600	7.400	3	3	2	1	1	7.200	8.073
4	4	1	4.967	6.667	3	3	3	2	6.727	8.015	3	3	2	2	1	7.591	8.576
4	4	2	5.455	7.036	3	3	3	3	7.000	8.538	3	3	2	2	2	7.910	9.115
4	4	3	5.598	7.144	4	1	1	1	–	–	3	3	3	1	1	7.576	8.424
4	4	4	5.692	7.654	4	2	1	1	5.391	–	3	3	3	2	1	7.769	9.051
5	2	1	5.000	–	4	2	2	1	6.133	7.000	3	3	3	2	2	8.044	9.505

(Contd.)

K = 3					K = 4						K = 5						
Sample sizes			α = 0.05	α = 0.01	Sample Sizes				α = 0.05	α = 0.01	Sample sizes					α = 0.05	α = 0.01
5	2	2	5.160	6.533	4	2	2	2	6.545	7.391	3	3	3	3	1	8.000	9.451
5	3	1	4.960	–	4	3	1	1	6.178	7.067	3	3	3	3	2	8.200	9.876
5	3	2	5.251	6.909	4	3	2	1	6.309	7.455	3	3	3	3	3	8.333	10.20
5	3	3	5.648	7.079	4	3	2	2	6.621	7.871							
5	4	1	4.985	6.955	4	3	3	1	6.545	7.758							
5	4	2	5.273	7.205	4	3	3	2	6.795	8.333							
5	4	3	5.656	7.445	4	3	3	3	6.984	8.659							
5	4	4	5.657	7.760	4	4	1	1	5.945	7.909							
5	5	1	5.127	7.309	4	4	2	1	6.386	7.909							
5	5	2	5.338	7.338	4	4	2	2	6.731	8.346							
5	5	3	5.705	7.578	4	4	3	1	6.635	8.231							
5	5	4	5.666	7.823	4	4	3	2	6.874	8.621							
5	5	5	5.780	8.000	4	4	3	3	7.038	8.876							
6	1	1	–	–	4	4	4	1	6.725	8.588							
6	2	1	4.822	–	4	4	4	2	6.957	8.871							
6	2	2	5.345	6.655	4	4	4	3	7.142	9.075							
6	3	1	4.855	6.873	4	4	4	4	7.235	9.287							
6	3	2	5.348	6.970													
6	3	3	5.615	7.410													
6	4	1	4.947	7.106													
6	4	2	5.340	7.340													
6	4	3	5.610	7.50													
6	4	4	5.681	7.795													
6	5	1	4.990	7.182													
6	5	2	5.338	7.376													
6	5	3	5.602	7.590													
6	5	4	5.661	7.936													
6	5	5	5.729	8.028													
6	6	1	4.945	7.121													
6	6	2	5.410	7.467													
6	6	3	5.625	7.725													
6	6	4	5.725	8.000													
6	6	5	5.765	8.124													
6	6	6	5.801	8.222													
7	7	7	5.819	8.378													
8	8	8	5.805	8.465													

Table 9: Cumulative Binomial Probabilities: $P(x \le r_0)$, $X \sim B(n, p)$

n	r_0	.10	.25	.40	.50
1	0	.9000	.7500	.6000	.5000
	1	1.0000	1.0000	1.0000	1.0000
2	0	.8100	.5625	.3600	.2500
	1	.9900	.9375	.8400	.7500
	2	1.0000	1.0000	1.0000	1.0000
5	0	.5905	.2373	.0778	.0313
	1	.9185	.6328	.3370	.1875
	2	.9914	.8965	.6826	.5000
	3	.9995	.9844	.9130	.8125
	4	.9999	.9990	.9898	.9687
	5	1.0000	1.0000	1.0000	1.0000
10	0	.3487	.0563	.0060	.0010
	1	.7361	.2440	.0463	.0108
	2	.9298	.5256	.1672	.0547
	3	.9872	.7759	.3822	.1719
	4	.9984	.9219	.6330	.3770
	5	.9999	.9803	.8337	.6230
	6	1.0000	.9965	.9452	.8281
	7	1.0000	.9996	.9877	.9453
	8	1.0000	1.0000	.9983	.9892
	9	1.0000	1.0000	.9999	.9990
	10	1.0000	1.0000	1.0000	1.0000
12	0	.2824	.0317	.0022	.0002
	1	.6590	.1584	.0196	.0031
	2	.8891	.3907	.0835	.0192
	3	.9740	.6488	.2254	.0729
	4	.9963	.8424	.4382	.1937
	5	.9999	.9456	.6652	.3871
	6	1.0000	.9857	.8418	.6127
	7	1.0000	.9972	.9427	.8064
	8	1.0000	.9996	.9847	.9269
	9	1.0000	1.0000	.9972	.9806
	10	1.0000	1.0000	.9997	.9977
	11	1.0000	1.0000	1.0000	1.0000
	12	1.0000	1.0000	1.0000	1.0000

(Contd.)

n	x	.10	.25	.40	.50
20	0	.1216	.0032	.0000	.0000
	1	.3917	.0243	.0005	.0000
	2	.6768	.0912	.0036	.0002
	3	.8669	.2251	.0159	.0013
	4	.9567	.4148	.0509	.0059
	5	.9886	.6171	.1255	.0207
	6	.9975	.7857	.2499	.0577
	7	.9995	.8981	.4158	.1316
	8	.9999	.9590	.5955	.2517
	9	1.0000	.9861	.7552	.4119
	10	1.0000	.9960	.8723	.5881
	11	1.0000	.9990	.9433	.7483
	12	1.0000	.9998	.9788	.8684
	13	1.0000	1.0000	.9934	.9423
	14	1.0000	1.0000	.9983	.9793
	15	1.0000	1.0000	.9996	.9941
	16	1.0000	1.0000	1.0000	.9987
	17	1.0000	1.0000	1.0000	.9998
	18	1.0000	1.0000	1.0000	1.0000
	19	1.0000	1.0000	1.0000	1.0000
	20	1.0000	1.0000	1.0000	1.0000

Table 10: Selected Critical Values of *S* in the Kendall's Coefficient of Concordance

Values at 5% level of significance

k	N					Some additional values for $N = 3$	
	3	4	5	6	7	k	s
3			64.4	103.9	157.3	9	54.0
4		49.5	88.4	143.3	217.0	12	71.9
5		62.6	112.3	182.4	276.2	14	83.8
6		75.7	136.1	221.4	335.2	16	95.8
8	48.1	101.7	183.7	299.0	453.1	18	107.7
10	60.0	127.8	231.2	376.7	571.0		
15	89.8	192.9	349.8	570.5	864.9		
20	119.7	258.0	468.5	764.4	1158.7		

Values at 1% level of significance

3			75.6	122.8	185.6	9	75.9
4		61.4	109.3	176.2	265.0	12	103.5
5		80.5	142.8	229.4	343.8	14	121.9
6		99.5	176.1	282.4	422.6	16	140.2
8	66.8	137.4	242.7	388.3	579.9	18	158.6
10	85.1	175.3	309.1	494.0	737.0		
15	131.0	269.8	475.2	758.2	1129.5		
20	177.0	364.2	641.2	1022.2	1521.9		

Table 11: Table Showing Critical Values of *A*-Statistic for any Given Value of *n* – 1, Corresponding to Various Levels of Probability

(*A* is significant at a given level if it is ≤ the value shown in the table)

$n-1$	Level of significance for one-tailed test				
	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test				
	.10	.05	.02	.01	.001
1	2	3	4	5	6
1	0.5125	0.5031	0.50049	0.50012	0.5000012
2	0.412	0.369	0.347	0.340	0.334
3	0.385	0.324	0.286	0.272	0.254
4	0.376	0.304	0.257	0.238	0.211
5	0.372	0.293	0.240	0.218	0.184
6	0.370	0.286	0.230	0.205	0.167
7	0.369	0.281	0.222	0.196	0.155
8	0.368	0.278	0.217	0.190	0.146
9	0.368	0.276	0.213	0.185	0.139
10	0.368	0.274	0.210	0.181	0.134
11	0.368	0.273	0.207	0.178	0.130
12	0.368	0.271	0.205	0.176	0.126
13	0.368	0.270	0.204	0.174	0.124
14	0.368	0.270	0.202	0.172	0.121
15	0.368	0.269	0.201	0.170	0.119
16	0.368	0.268	0.200	0.169	0.117
17	0.368	0.268	0.199	0.168	0.116
18	0.368	0.267	0.198	0.167	0.114
19	0.368	0.267	0.197	0.166	0.113

$n-1^*$	Level of significance for one-tailed test				
	.05	.025	.01	.005	.0005
	Level of significance for two-tailed test				
	.10	.05	.02	.01	.001
1	2	3	4	5	6
20	0.368	0.266	0.197	0.165	0.112
21	0.368	0.266	0.196	0.165	0.111
22	0.368	0.266	0.196	0.164	0.110
23	0.368	0.266	0.195	0.163	0.109
24	0.368	0.265	0.195	0.163	0.108
25	0.368	0.265	0.194	0.162	0.108
26	0.368	0.265	0.194	0.162	0.107
27	0.368	0.265	0.193	0.161	0.107
28	0.368	0.265	0.193	0.161	0.106
29	0.368	0.264	0.193	0.161	0.106
30	0.368	0.264	0.193	0.160	0.105
40	0.368	0.263	0.191	0.158	0.102
60	0.369	0.262	0.189	0.155	0.099
120	0.369	0.261	0.187	0.153	0.095
∞	0.370	0.260	0.185	0.151	0.092

* n = number of pairs

Source: *The Brit. J. Psychol*, Volume XLVI, 1955, p. 226.

Selected References and Recommended Readings

1. Ackoff, Russell L., *The Design of Social Research,* Chicago: University of Chicago Press, 1961.
2. Ackoff, Russell L., *Scientific Method,* New York: John Wiley & Sons, 1962.
3. Allen, T. Harrell, *New Methods in Social Science Research,* New York: Praeger Publishers, 1978.
4. Anderson, H.H., and Anderson, G.L., *An Introduction to Projective Techniques and Other Devices for Understanding the Dynamics of Human Behaviour*, New York: Prentice Hall, 1951.
5. Anderson, T.W., *An Introduction to Multivariate Analysis,* New York: John Wiley & Sons, 1958.
6. Bailey, Kenneth D., *Methods of Social Research,* New York, 1978.
7. Baker, R.P., and Howell, A.C., *The Preparation of Reports,* New York: Ronald Press, 1938.
8. Bartee, T.C., *Digital Computer Fundamentals,* 5th Ed., McGraw-Hill, International Book Co., 1981.
9. Barzun, Jacques, and Graff, Henery, F., *The Modern Researcher,* rev. ed., New York: Harcourt, Brace & World, Inc., 1970.
10. Bell, J.E., *Projective Techniques: A Dynamic Approach to the Study of Personality,* New York: Longmans Green, 1948.
11. Bellenger, Danny N., and Greenberg, Barnett A., *Marketing Research—A Management Information Approach,* Homewood, Illinois: Richard D. Irwin, Inc., 1978.
12. Berdie, Douglas R., and Anderson, John F., *Questionnaires: Design and Use,* Metuchen N.J.: The Scarecrow Press, Inc., 1974.
13. Berelson, Bernard, *Content Analysis in Communication Research,* New York: Free Press, 1952.
14. Berenson, Conard, and Colton, Raymond, *Research and Report Writing for Business and Economics,* New York: Random House, 1971.
15. Best, John W., and Kahn, James V., *Research in Education*, 5th ed., New Delhi: Prentice-Hall of India Pvt. Ltd., 1986.
16. Bhattacharya, Srinibas, *Psychometrics & Behavioural Research,* New Delhi: Sterling Publishers Pvt. Ltd., 1972.
17. Boot, John C.G., and Cox, Edwin B., *Statistical Analysis for Managerial Decisions,* 2nd ed., New Delhi: McGraw-Hill Publishing Co. Ltd., (International Student Edition), 1979.
18. Bowley, A.L., *Elements of Statistics,* 6th ed., London: P.S. King and Staples Ltd., 1937.

19. Burgess, Ernest W., "Research Methods in Sociology" in Georges Gurvitch and W.E. Moore (Ed *Twentieth Century Sociology,* New York: New York Philosophical Library, 1949.

20. Chance, William A., *Statistical Methods for Decision Making,* Bombay: D.B. Taraporevala Sons & C Pvt. Ltd., 1975.

21. Chaturvedi, J.C., *Mathematical Statistics,* Agra: Nok Jhonk Karyalaya, 1953.

22. Chou, Ya-Lun, *Statistical Analysis with Business and Economic Applications,* 2nd ed., New York: Hol Rinehart & Winston, 1974.

23. Clover, Vernon T., and Balsley, Howard L., *Business Research Methods,* Columbus, O.: Grid, Inc., 1974

24. Cochran, W.G., *Sampling Techniques,* 2nd ed., New York: John Wiley & Sons., 1963.

25. Cooley, William W., and Lohnes, Paul R., *Multivariate Data Analysis*, New York: John Wiley & Sons 1971.

26. Croxton, F.E., Cowden, D.J., and Klein, S., *Applied General Statistics,* 3rd ed., New Delhi: Prentice-Hall India Pvt. Ltd., 1975.

27. Dass, S.L., *Personality Assessment Through Projective Movie Pictures*, New Delhi: S. Chand & C (Pvt.) Ltd., 1974.

28. Davis, G.B., *Introduction to Computers*, 3rd ed., McGraw-Hill International Book Co., 1981.

29. Deming, W. Edwards. *Sample Design in Business Research,* New York: John Wiley & Sons., Inc., 1960

30. Dennis, Child, *The Essentials of Factor Analysis,* New York: Holt, Rinehart and Winston, 1973.

31. Denzin, Norman, *The Research Act,* Chicago: Aldine, 1973.

32. Edwards, Allen, *Statistical Methods,* 2nd ed., New York: Holt, Rinehart & Winston, 1967.

33. Edwards, Allen L., *Techniques of Attitude Scale Construction,* New York: Appleton-Century-Crofts 1957.

34. Emory, C. William, *Business Research Methods*, Illinois: Richard D. Irwin, Inc., Homewood, 1976.

35. Ferber, Robert (Ed.), *Handbook of Marketing Research,* New York: McGraw-Hill, Inc., 1948.

36. Ferber, R., and Verdoorn P.J., *Research Methods in Economics and Business,* New York: The Macmilla Company, 1962.

37. Ferguson, George A., *Statistical Analysis in Psychology and Education*, 4th ed., New York: McGraw-Hil Book Co., Inc., 1959.

38. Festinger, Leon and Katz, Daniel (Eds.), *Research Methods in the Behavioral Sciences,* New Delhi Amerind Publishing Co. Pvt. Ltd., Fourth Indian Reprint, 1976.

39. Fiebleman, J.K., *Scientific Method*, Netherlands: Martinus Nijhoff, The Hague, 1972.

40. Fisher, R.A., *Statistical Methods for Research Workers,* 13th ed., New York: Hafner Publishing Co., 1958

41. Fisher, R.A., *The Design of Experiments,* 7th rev. ed., New York: Hafner Publishing Co., 1960.

42. Fox, James Harold, *Criteria of Good Research,* Phi Delta Kappa, Vol. 39 (March 1958).

43. Freedman, P., *The Principles of Scientific Research,* 2nd ed., New York: Pergamon Press, 1960.

44. Fruchter, Benjamin, *Introduction to Factor Analysis,* Princeton, N.J.: D.Van Nostrand, 1954.

45. Gatner, Elliot S.M., and Cordasco, Francesco, *Research and Report Writing*, New York: Barnes & Noble Inc., 1956.

46. Gaum, Carl G., Graves, Harod F., and Hoffman, Lyne, S.S., *Report Writing*, 3rd ed., New York: Prentice-Hall 1950.

47. Ghosh, B.N., *Scientific Methods and Social Research,* New Delhi: Sterling Publishers Pvt. Ltd., 1982.

48. Gibbons, J.D., *Nonparametric Statistical Inference,* Tokyo: McGraw-Hill Kogakusha Ltd., (International Student Edition), 1971.

49. Giles, G.B., *Marketing,* 2nd ed., London: Macdonald & Evans Ltd., 1974.
50. Glock, Charles Y., *Survey Research in the Social Sciences,* New York: Russell Sage Foundation, 1967.
51. Godfrey, Arthur, *Quantitative Methods for Managers,* London: Edward Arnold (Publishers) Ltd., 1977.
52. Good, Carter V., and Douglas, E. Scates, *Methods of Research—Educational, Psychological, Sociological,* New York: Appleton-Century-Crofts, Inc., 1954.
53. Goode, William J., and Hatt, Paul K., *Methods in Social Research,* New York: McGraw-Hill, 1952.
54. Gopal, M.H., *An Introduction to Research Procedure in Social Sciences,* Bombay: Asia Publishing House, 1964.
55. Gopal, M.H., *Research Reporting in Social Sciences,* Dharwar: Karnatak University, 1965.
56. Gorden, Raymond L., *Interviewing: Strategy, Techniques and Tactics,* rev. ed., Homewood, Ill.: Dorsey Press, 1975.
57. Green, Paul E., *Analyzing Multivariate Data,* Hinsdale, Ill.: Dryden Press, 1978.
58. Green, Paul E., and Carmone, F.J., *Multidimensional Scaling in Marketing Analysis,* Boston: Allyn & Bacon, Inc., 1970.
59. Guilford, J.P., *Psychometric Methods,* New York: McGraw Hill, Inc., 1954.
60. Harnett, Donald L., and Murphy, James L., *Introductory Statistical Analysis,* Philippines: Addison-Wesley Publishing Co., Inc., 1975.
61. Hillway, T., *Introduction to Research,* 2nd ed., Boston: Houghton Mifflin, 1964.
62. Hollander, Myles, and Wolfe, Douglas A., *Nonparametric Statistical Methods,* New York: John Wiley, 1973.
63. Hunt, R., and Shelley, J., *Computers and Common Sense,* 3rd ed., New Delhi: Prentice-Hall of India Ltd., 1984.
64. Hyman, Herbert H., et al., *Interviewing in Social Research,* Chicago: University of Chicago Press, 1975.
65. John W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches,* New Delhi: SAGE Publications.
66. John W. Creswell, Cheryl N. Poth, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, New Delhi: SAGE Publications.
67. John, Peter W.M., *Statistical Design and Analysis of Experiments,* New York: The Macmillan Co., 1971.
68. Johnson, Ellen, *The Research Report: A Guide for the Beginner,* New York: Ronald Press, 1951.
69. Johnson, Rodney D., and Siskin, Bernard R., *Quantitative Techniques for Business Decisions,* New Delhi: Prentice-Hall of India Pvt. Ltd., 1977.
70. Kahn, Robert L. and Cannell, Charles F., *The Dynamics of Interviewing*, New York: John Wiley & Sons, 1957.
71. Karson, Marvin J., *Multivariate Statistical Methods*, Ames, Iowa: The Iowa State University Press, 1982.
72. Kendall, M.G., *A Course in Multivariate Analysis,* London, Griffin, 1961.
73. Kerlinger, Fred N. and Pedhazur, Elazar J., *Multiple Regression in Behavioral Research,* New York: Holt, Rinehart and Winston, 1973.
74. Kerlinger, Fred N., *Foundations of Behavioral Research,* 2nd ed., New York: Holt, Reinhart and Winston, 1973.
75. Kish, Leslie., *Survey Sampling,* New York: John Wiley & Sons, Inc., 1965.
76. Kothari, C.R., *Quantitative Techniques,* 2nd ed., New Delhi: Vikas Publishing House Pvt. Ltd., 1984.
77. Lastrucci, Carles L., *The Scientific Approach: Basic Principles of the Scientific Method,* Cambridge, Mass.: Schenkman Publishing Co., Inc., 1967.

78. Lazersfeld, Paul F., "*Evidence and Inference in Social Research,*" in David Lerher, *Evidence and Inferen*
Glencoe: The Free Press, 1950.

79. Leonard Schatzman, and Anselm L. Strauss, *Field Research,* New Jersey: Prentice-Hall Inc., 1973.

80. Levin, Richard I., *Statistics for Management,* New Delhi: Prentice-Hall of India Pvt. Ltd., 1979.

81. Levine, S., and Elzey, Freeman F., *A Programmed Introduction to Research,* California: Wodswo
Publishing Co., 1968.

82. Maranell, Gary M. (Ed.), *Scaling: A Source Book for Behavioral Scientists,* Chicago: Aldine, 1974.

83. Maxwell, Albert E., *Analyzing Qualitative Data,* New York: John Wiley & Sons, 1961.

84. Meadows, R., and Parsons, A.J., *Microprocessors: Essentials, Components and Systems,* Pitman, 1983.

85. Meir, Robert C., Newell, William T., and Dazier, Harold L., *Simulation in Business and Economi*
Englewood Cliffs, N.J: Prentice Hall, Inc., 1969.

86. Michael Crotty, *The Foundations of Social Research: Meaning and Perspective in the Researc*
New Delhi: SAGE Publications.

87. Miller, Delbert C., *Handbook of Research Design & Social Measurement,* 3rd ed., New York: Dav
Mckay Company, Inc., 1977.

88. Moroney, M.J., *Facts from Figures,* Baltimore: Penguin Books, 1956.

89. Morrison, Donald F., *Multivariate Statistical Methods,* New York: McGraw-Hill, 1967.

90. Nagel, Stuart S., and Neef, Marian, *Policy Analysis in Social Science Research,* London: Sage Publicatio
1979.

91. Nie, N.H., Bent, D.H., and Hull, C.H., *Statistical Package for the Social Sciences,* New York: McGra
Hill, 1970.

92. Noether, G.E., *Elements of Nonparametric Statistics,* New York: John Wiley & Sons, Inc., 1967.

93. Nunnally, Jum C., *Psychometric Theory,* 2nd ed., New York: McGraw-Hill, 1978.

94. Odum, H.W., and Jocher, Katharine, *An Introduction to Social Research,* New York: Henry Holt and C
1929.

95. Oppenheim, A.N., *Questionnaire Design and Attitude Measurement,* New York: Basic Books, 1966.

96. Ostle, Bernard, and Mensing, Richard W., *Statistics in Research,* 3rd ed., Ames Iowa: The Iowa Sta
University Press, 1975.

97. Payne, Stanley, *The Art of Asking Questions,* Princeton: Princeton University Press, 1951.

98. Pearson, Karl, *The Grammar of Science,* New York: Meridian Books, Inc., 1957.

99. Phillips, Bernard S., *Social Research, Strategy and Tactics,* 2nd ed., New York: The Macmilla
Company, 1971.

100. Piaget, Jean, *Main Trends in Interdisciplinary Research,* London: George Allen and Unwin Ltd., 1973

101. Popper, Karl R., *The Logic of Scientific Discovery,* New York: Basic Books, 1959.

102. Rajaraman, V., *Fundamentals of Computers,* New Delhi: Prentice-Hall of India Pvt. Ltd., 1985.

103. Ramchandran, P., *Training in Research Methodology in Social Sciences in India,* New Delh
ICSSR 1971.

104. Redman, L.V., and Mory, A.V.H., *The Romance of Research,* 1923.

105. Roscoe, John T., *Fundamental Research Statistics for the Behavioral Sciences,* New York: Holt, Rineha
and Winston, Inc., 1969.

106. Runyon, Richard P., *Inferential Statistics,* Philippines: Addison-Wesley Publishing Company, Inc., 197

107. Sadhu, A.N., and Singh, Amarjit, *Research Methodology in Social Sciences,* Bombay: Himalaya Publishin
House, 1980.

108 Seboyar, G.E., *Manual for Report and Thesis Writing,* New York: F.S. Crofts & Co., 1929.

9. Selltiz, Claire: Jahoda, Marie, Deutsch, Morton, and Cook, Stuart W., *Research Methods in Social Relations,* rev. ed., New York: Holt, Rinehart and Winston, Inc., 1959.

10. Sharan B. Merriam, Elizabeth J. Tisdell, *Qualitative Research: A Guide to Design and Implementation*, New Delhi: Wiley.

11. Sharma, B.A.V., et al., *Research Methods in Social Sciences,* New Delhi: Sterling Publishers Pvt. Ltd., 1983.

12. Sharma, H.D., and Mukherji, S.P., *Research in Economics and Commerce, Methodology and Surveys,* Varanasi: Indian Biographic Centre, 1976.

13. Siegel, S., *Nonparametric Statistics for the Behavioral Sciences,* New York: McGraw-Hill Publishing Co., Inc., 1956.

14. Subramanian, N. *Introduction to Computers*, New Delhi: Tata McGraw-Hill Publishing Co. Ltd., 1986.

15. Summers, Gene F. (Ed.), *Attitude Measurement,* Chicago: Rand McNally & Co., 1970.

16. Takeuchi, K., Yanai, H., and Mukherjee, B.N., *The Foundations of Multivariate Analysis*, New Delhi: Wiley Eastern Ltd., 1982.

17. Tandon, B.C., *Research Methodology in Social Sciences,* Allahabad: Chaitanya Publishing House, 1979.

18. Thorndike, Robert L., and Hagen, Elizabeth P., *Measurement and Evaluation in Psychology and Education,* 4th ed., New York: John Wiley & Sons, 1977.

19. Thurstone, L.L., *The Measurement of Values,* Chicago: University of Chicago Press, 1959.

20. Torgerson, W., *Theory and Methods of Scaling,* New York: John Wiley & Sons, 1958.

21. Travers, Robert M.W., *An Introduction to Educational Research,* 4th ed., New York: Macmillan Publishing Co., Inc., 1978.

22. Tryon, R.C., and Bailey, D.E., *Cluster Analysis*, New York: McGraw-Hill, 1970.

23. Ullman, Neil R., *Elementary Statistics,* New York: John Wiley & Sons, Inc., 1978.

24. Wayne C. Booth, Gregory G. Colomb, Joseph M. Williams, *The Craft of Research, 3rd ed.,* Chicago: University of Chicago Press.

25. Whitney, F.L., *The Elements of Research,* 3rd ed., New York: Prentice-Hall, 1950.

26. Wilkinson, T.S., and Bhandarkar, P.L., *Methodology and Techniques of Social Research,* Bombay: Himalaya Publishing House, 1979.

27. William Trochim, James P. Donnelly, *The Research Methods Knowledge Base*, New Delhi: Cengage Learning.

28. Willemsen, Eleanor Walker, *Understanding Statistical Reasoning*, San Francisco: W.H. Freeman and Company, 1974.

29. Yamane, T., *Statistics: An Introductory Analysis,* 3rd ed., New York: Harper and Row, 1973.

30. Young, Pauline V., *Scientific Social Surveys and Research,* 3rd ed., New York: Prentice-Hall, 1960.

31. Johnson, Richards A. and Wichern, Dean W., Applied Multivariate Statistical Analysis (5th Ed.), Prentice Hall, Englewood Cliffs, New Jersey.

Index

After-only with control design, 39
Analysis of covariance (ANOCOVA), 281, 289
Analysis of data, 9, 12, 17, 30, 52
  classification, 114
  coding, 17, 114
  editing, 17, 114
  tabulation, 17, 114, 119
  types of analysis, 126
Analysis of variance (ANOVA), 261
  one-way, 262, 271
  two-way, 262, 271
ANOVA in Latin square design, 277, 285
Analytical research, 2
Applied research, 2
Area sampling, 15, 62

Before-and-after design:
  with control, 13
  without control, 13
Beta error, 346
Bibliography, 18, 410
Bipolar factor, 365

Canonical correlation analysis, 414
Case-study method, 4
Census survey, 52
Central limit theorem, 150, 154
Centroid method of factor analysis, 362
Chi-square test, 142
Classification, 114, 116
  according to attributes, 116
  according to class-intervals, 116
Cluster analysis, 358
Cluster sampling, 15, 62
Coding, 17, 114
Coefficient of—
  association, 128, 141
  contingency, 128, 243
  correlation, 33, 70
  reproducibility, 88
  skewness, 136
  standard deviation, 128, 134
  variation, 128, 134
Collection of primary data, 91
  interview method, 92
    depth interviews, 106
    personal interviews, 92
    telephone interviews, 95
  observation method, 91
  questionnaires, 96
    through schedules, 100
Completely randomized design (C.R. Design), 39
Composite standard method, 77
Computer: 127, 170
  digital, 382
Confidence level, 53, 157
Confounded relationship, 51
Construct validity, 70
Consumer panels, 103
Content analysis, 107
Content validity, 70
Contingency table, 143, 236
Continuous variable, 32
Control, 2, 32

- Control groups, 33, 42
- Convenience sampling, 13, 14
- Correlation analysis, 126, 127
- Criteria of good research, 21
- Criterion-related validity, 70
- Cross tabulation, 120, 236

- Deliberate sampling, 14, 55
- Depth interviews, 106, 107
- Descriptive research, 2, 35
- Descriptive research studies, 35, 90
- Discrete variable, 32
- Distribution—free tests, 290
  - (*See* non-parametric tests)

- Editing, 115
  - central editing, 115
  - field editing, 115
- Empirical research, 3, 20
- Estimation, 56, 156
  - interval estimate, 154, 157
  - point estimate, 154
- Experience survey, 26, 35
- Experiment, 34
- Experimental designs, 38, 39
  - formal, 39, 40
    - C.R. design, 39, 41
    - Factorial design, 46, 48
    - L.S. design, 45, 46
    - R.B. design, 44
  - informal, 39, 41
- Experimental group, 33, 282
- Experimental research, 32, 89
- Experimental unit, 34
- Exploratory research studies, 34
- Ex-post facto research, 68
- Extraneous variable, 32, 42

- *F*-distribution, 152, 184
- *F*-test, 323, 334
- Factor analysis, 358, 366
  - factor loadings, 359, 366
  - factor scores, 359
- Factorial designs, 46, 51
- Fundamental research, 2

- Holtzman Inkblot test (H.I.T), 105
- Hypothesis:
  - alternative, 180, 181
  - characteristics of, 180
  - meaning, 179
  - null, 180, 181
- Hypothesis testing, 183
- Hypothesis testing, research studies, 2, 38

- Index numbers, 128, 143
- Interaction effect, 47, 48, 261
- Interaction variation, 274
- Interval,
  - estimates, 154
  - scales, 67, 68
- Interview method, 92, 93, 94, 109
  - depth interviews, 106
  - personal interviews, 92
    - clinical, 93
    - focussed, 93
    - non-directive, 93
    - structured, 93
    - unstructured, 92
  - telephone interviews, 95

- Judgement sampling, 14

- Kruskal-Wallis test, 303, 310

- Latent structure analysis, 415
- Latin square design (L.S. Design), 39
- Law of comparative judgement, 76
- Level of significance, 175
- Likert-type scales, 80, 81
- Literature survey, 12
- Longitudinal research, 3

- Maximum likelihood method, 358
- Measures of central tendency, 128, 129
  - geometric mean, 131
  - harmonic mean, 131, 132
  - mean, 129, 130
  - median, 129, 130
  - mode, 129
- Measures of dispersion, 128, 129
  - mean deviation, 132
  - range, 132
  - standard deviation, 133
- Measures of relationship, 128
- Measures of skewness, 128, 136
  - Kurtosis, 128
- Measurement in research, 66
  - interval data, 67

- nominal data, 67
- ordinal data, 67
- ratio data, 67

Measurement scales, 67
- interval scale, 67
- nominal scale, 67
- ordinal scale, 67
- ratio scale, 67

Motivation research, 3
Multicollinearity, 33
Multi stage sampling, 15, 62
Multivariate analysis techniques, 410
- factor analysis, 357
- multidimensional scaling, 85, 410
  - metric approach, 86
  - non-metric approach, 89
- multiple correlation, 128
- multiple discriminant analysis, 127, 358
- multiple regression, 349
- multivariate analysis of variance, 127, 411

Nominal scale, 68
Non-parametric tests, 235, 290, 309
- Chi-square test, 142, 237

Null-hypothesis, 268
- One sample runs test, 301, 302
- Signed rank test, 294
- Sign tests, 291
- Spearman's rank correlation, 305
- Rank sum tests: 294
  - H-test, 303
  - U-test, 297

Observation method, 91, 94
- controlled observation, 92
- disguised observation, 92
- participant observation, 92
- structured observation, 91
- unstructured observation, 92

Paired *t*-test, 217
Pantry audits, 102
Parametric tests, 235, 291
Partial correlation, 128
Path analysis, 410, 412
Pilot survey, 25
Post factum interpretation, 417
Precision, 109, 149
Primary data, 89
Principal-components method, 358, 366
Principles of experimental designs, 38, 44
- local control, 39
- randomization, 41
- replication, 41

Projective techniques, 104

Q-type factor analysis, 371
Questionnaire, 97, 98
Quota sampling, 15, 56

Randomized Block Design (R.B. Design), 44
Random sampling, 14, 56
Ranking scales, 74
Rating scales, 79
- graphic rating, 79
- itemized rating, 79

Regression, 128
Report writing, 423
Research:
- meaning of, 8
- motivation in, 3
- objectives of, 2
- significance of 5
- types of, 2

Research approaches, 4
Research design, 12, 13, 29
Research hypothesis, 33
Research methods, 6
Research methodology, 6
Research problem, 22, 23
Research process, 9, 22
Rosenzweig test, 105
R-type factor analysis, 371

Sample design, 9, 13, 52
- non-probability sampling, 55
- probability sampling, 56
- random sampling, 56, 58
- stratified sampling, 13, 14

Sample size, 54, 161
Sample survey, 52
Sampling, 147
- design, 37
- distribution, 149
- meaning, 148
- need, 147
- theory, 170

Sampling error, 54, 55, 149
Sampling frame, 51, 53
Sampling unit, 55
Sandler's *A*-test, 230
Scale construction techniques, 75
  arbitrary scales, 75
  cumulative scales, 75
  differential scales, 82
  factor scales, 75
  multidimensional scaling, 75
  semantic differential scales, 82
  summated scales, 79
Scaling, 73
Scientific method, 8
Secondary data, 89
  collection of, 107
Sequential sampling, 15, 64
Significance level, 163
Sociogram, 106
Sociometry, 106
Source list, 53
Sources of error in measurement, 72
Standard error, 152
Stores audit, 102
Stratified sampling, 13
Student's *t*-test, 231
Survey design, 31, 37
Systematic sampling, 13, 14, 58

Tabulation, 119
  accepted principles of, 120
Tally sheet, 120
Technique of developing,
  measurement tools; 72
Testing of Hypotheses,
  basic concepts concerning, 180
Tests of hypotheses
  non-parametric tests, 309
  parametric tests, 230
Tests of sound measurement, 71
Thematic apperception test (T.A.T.), 105
Thurstone-type scales, 81
Time-series analysis, 145
Tomkins-Horn picture arrangement test, 106
t-test, 209
Two-tailed and one-tailed test, 229
Type I and Type II errors, 181
Types of analysis, 126
  bivariate, 126
  causal, 126
  correlation, 126
  descriptive, 126
  inferential, 126
  multivariate, 126
  unidimensional, 126

Variable, 32, 33, 281
  continuous, 32
  criterion, 382
  dependent, 32, 312
  discrete, 32
  dummy, 341
  explanatory, 312
  extraneous, 32
  independent, 32
  latent, 357
  observable, 357
  pseudo, 357
Varimax rotation, 376

Warranty cards, 102

Yate's correction, 242
Yule's coefficient of association, 128, 141

Z-test, 213

Purchased:
Approval : MLK
Gratis : 038805
Acc No :
Price :
"Narayan Rao Melgiri
National Law Library"
Bangalore
h 3991

300.72
Cl.No. KOT-2
Author Kothari

300.72
KOT-2

38805

Primary Research
Data Validation
360° Research
Competitors Sources

RESEARCH METHODOLOGY
METHODS AND TECHNIQUES

Contents:

- Research Methodology: An Introduction
- Defining the Research Problem
- Research Design
- Design of Sample Surveys
- Measurement and Scaling
- Data Collection
- Data Preparation
- Descriptive Statistics
- Sampling and Statistical Inference
- Testing of Hypothesis
- Chi-Square Tests
- Analysis of Variance
- Other Nonparametric Methods
- Linear Regression Analysis
- Factor Analysis
- Discriminant Analysis
- Cluster Analysis
- Other Multivariate Techniques
- Interpretation and Report Writing

This fourth multi colour edition has been made better equipped to make the learning more systematic. More details are added to old chapters keeping in mind the need of advanced data analysis in social science research. The usage of popular statistical package SPSS has also been shown through many examples.

It is important to adopt an appropriate methodology in quality research. The present book provides the basic and advanced level discussion on research methodology so that the researchers may become familiar with the art of using research methods and techniques.

The book is intended to serve as a textbook for social science students at graduate and postgraduate level. The book can be taken as a well-organised guide for researchers in social science and other areas.

Dr. C R Kothari was Principal, College of Commerce and Associate Professor in the Department of Economic Administration and Financial Management, University of Rajasthan, Jaipur, Rajasthan (India).

Dr. Gaurav Garg is Associate Professor in Decision Sciences Area, Indian Institute of Management, Lucknow, Uttar Pradesh (India).

MRP: ₹ 399.00

ISBN: 978-93-86649-22-5

9 789386 649225

NEW AGE
TECHNO
PRESS

An Imprint of

NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS
LONDON • NEW DELHI • NAIROBI
Bangalore • Chennai • Cochin • Guwahati • Hyderabad • Kolkata • Lucknow • Mumbai
Visit us at **www.newagepublishers.com**